Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra počítačových systémů

Bakalářská práce

Digitalizace kartoték

Jan Anděl

Vedoucí práce: Ing. Radek Dobiáš, Ph.D.

16. prosince 2014

Poděkování

Mé poděkování patří Ing. Radkovi Dobiášovi, Ph.D. za odborné vedení, trpělivost a ochotu, kterou mi v průběhu zpracování bakalářské práce věnoval.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 16. prosince 2014 .....

České vysoké učení technické v Praze
Fakulta informačních technologií
© 2014 Jan ANDĚL. Všechna práva vyhrazena.
Tato práce vznikla jako školní dílo na Českém vysokém učení technickém v Praze, Fakultě informačních technologií. Práce je chráněna právními předpisy a mezinárodními úmluvami o právu autorském a právech souvisejících s právem autorským. K jejímu užití, s výjimkou bezúplatných zákonných licencí, je nezbytný souhlas autora.

Odkaz na tuto práci

ANDĚL, Jan. *Digitalizace kartoték*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2014.

Abstrakt

Předmětem zájmu této práce jsou lístkové kartotéky a jejich možné způsoby digitalizace. Jsou popisovány teoretické postupy, jak s omezenými finančními i lidskými zdroji v co nejkratším čase provést oskenování kartotéky a vytěžení co možná nejvíce dat bez nutnosti přepisu. Zároveň je popsáno praktické řešení systémem RETROBI a poskytnut přehled existujících řešení digitalizovaných kartoték v České republice i v zahraničí. Je připojena část zabývající se automatickou instalací systému RETROBI a zajištěním bezpečnosti aplikačními firewally.

Klíčová slova Digitalizace, kartotéka, knihovní katalog, skenování, OPAC, RETROBI, aplikační firewall

Abstract

The objective of this bachelor thesis are different forms of card catalogues and potential methods of their digitalization. It describes theoretical approaches on how to maximise the time effectiveness of the digitalization and how to abstract as much information as possible without manual entry involvement, all with limited financial and human resources. The thesis also describes a practical reference solution by using RETROBI system and provides overview of currently available digitalization solutions in Czech republic and abroad. There is also a section which covers auto installation of RETROBI system and security measures managed by application firewalls.

Keywords Digitization, card catalog, library catalog, scanning, OPAC, RETROBI, application firewall

Obsah

Úvod **1**

1 Digitalizace kartoték v České republice **3**

2 Digitalizované kartotéky online, existující řešení **5**
2.1 Digitalizace skenováním a částečným přepisem 5
2.2 Databázové knižní katalogy 10
2.3 Situace v zahraničí 12

3 Obecný postup převodu lístkových kartoték **17**
3.1 Analyzování cílů, možností a prostředků 17
3.2 Stanovení cílů 19
3.3 Vybírání technologií 21
3.4 Příprava kartotéky 23
3.5 Skenování kartotéky 24
3.6 Vyvíjení aplikace 26
3.7 Vyvíjení databázové aplikace 28
3.8 Plnění aplikace 28

4 Systém RETROBI **29**
4.1 Retrospektivní bibliografie v ÚČL AV ČR, v. v. i. 29
4.2 Projekt RETROBI 30
4.3 Příprava kartotéky pro digitalizaci 31
4.4 Instalace aplikace RETROBI 35
4.5 Zálohování dat 38
4.6 RETROBI z pohledu obecné analýzy 39
4.7 RETROBI z pohledu uživatele 40

5 Bezpečnost aplikace RETROBI **43**
5.1 Funkce mod_security 45
5.2 Instalace a aktivace mod_security 45
5.3 Testování a výstup logu 47

Závěr **51**

Literatura **53**

A Seznam použitých termínů **57**

B Seznam použitých zkratek **59**

C Obsah přiloženého CD **61**

Seznam obrázků

2.1 Lístkový katalog/ Card catalogue, NPÚ 6
2.2 KATIF, NK ČR 7
2.3 Generální katalog, Moravská zemská knihovna 8
2.4 GEK, Vědecká knihovna v Olomouci 9
2.5 UCL Library Services, University College London 12
2.6 KatZoom, volba písmene, Universität Wien 13
2.7 KatZoom, volba indexu, Universität Wien 13
2.8 AK I, Universitätsbibliothek Greifswald 14
2.9 Digitalisierte Zettelkataloge der Sächsischen Landesbibliothek, Staats- und Universitätsbiliothek Dresden 15
2.10 Alphabetischer Katalog, Philosophische Fakultät der Universität Köln 16

3.1 Model převodu lístkových kartoték 18
3.2 Model stanovení cílů 21
3.3 Model vybírání technologií 22
3.4 Model přípravy kartotéky 24

4.1 Přehled indexů RBČL pro heslo „Vodák Jindřich“ 32
4.2 Systém RETROBI, ÚČL AV ČR v. v. i. 39

5.1 Ukázka probíhajícího útoku, mod_security v módu pasivního logování „DETECTION ONLY“ 48
5.2 Ukázka zamítnutí požadavku útočníka, mod_security v aktivním módu „ENABLED“ 49

Úvod

Podle údajů NIPOS, Centra informací a statistik kultury, bylo v roce 2011 v České republice 490 muzeí, galerií a památníků[33] a 5408 knihoven[34].

Bez výjimky každá z těchto institucí se musela někdy v minulosti, v době svého vzniku, rozhodnout, jakým způsobem bude inventarizovat svůj majetek, v jaké podobně a jaké o něm bude uchovávat informace.

Drtivá většina těchto institucí vznikla dávno před érou počítačů a sofistikovaných databází, a proto jsou dnes tyto informace zpravidla uloženy v lístkových kartotékách. Ty ukrývají někdy bezcenná, ale mnohdy nedocenitelná data, ke kterým má přístup pouze několik vyvolených z řad zaměstnanců dané instituce.

Na vlastníka katalogu působí dva protichůdné zájmy. Na jedné straně stojí nutnost existující vědomosti poskytnout co nejširšímu okruhu zájemců, zpřesňovat a aktualizovat již nabyté vědomosti, čistit je od chyb a nepřesností, uvádět je do souvislostí. Na druhé straně je zde potřeba jít dál, získávat nové vědomosti, zakládat nové projekty, stavět na již hotovém a zbytečně se nevracet.

Nabízejí se otázky, jak dlouho je možné starý systém používat bez inovace? Kdy je ještě rozumné i ekonomické rozvíjet původní systém starými prostředky i metodikou a kdy už je potřeba udělat krok kupředu a systém od základů přepracovat?

Lístkové kartotéky mají mnohaletou tradici, propracovaný systém třídění a tvoří důležitou vědomostní základnu mnoha kulturních i výzkumných institucí.

Tato práce pojednává o možnostech, jak lístkové kartotéky při maximálně efektivním využití omezených finančních prostředků převádět do digitální podoby.

Kapitola 1

Digitalizace kartoték v České republice

Otázku problematiky digitalizace kartoték začali řešit vlastníci kartoték již v počátcích nástupu cenově dostupné výpočetní techniky, datující se polovinou osmdesátých let.

V roce 1994 z iniciativy hlavní knihovnické autority České republiky, Národní knihovny, v rámci Národního programu retrospektivní konverze, začala vznikat technologie RETROKON[24] a Národní informační systém pro retrokonverzi (NRIS). Metodika a programové vybavení umožňující knihovnám skenování a publikování lístkových katalogů na Internetu.

I když tato technologie už v dobách svého vzniku zaznamenala mimořádný ohlas, s ohledem na nedostatek finančních prostředků byla testována jen na menších vzorcích a do plného provozu se začala dostávat až po několika letech.

V roce 2014 jsou tímto systémem zpřístupněny katalogy 13 knihoven[38] v České republice .

Nezávisle na projektu RETROKON vytváří Knihovna Akademie věd svůj vlastní systém Lístkový katalog[1], určený k prohlížení skenovaných kartoték souborných katalogů ústavů Akademie věd.

Objevují se i komerční subjekty s řešením vlastním nebo postaveným na zahraničních systémech.

V dnešní době proto existuje mnoho oddělených souhrnných katalogů založených na různých technologiích, s různým uživatelským komfortem a v různém stupni rozpracování.

Celkový přehled všech kartoték přístupných online bohužel není na žádném místě centrálně spravován.

Aktivita Národní knihovny, Národní program retrospektivní konverze, na svých stránkách `http://nprk.nkp.cz` prováděla sběr a publikaci informací o katalozích knihoven v ČR, pro které se plánuje, nebo již probíhá, retrospektivní konverze[39][22]. Výčet ale není zdaleka úplný.

Na rozdíl od databázových knižních katalogů, neexistuje pro digitalizaci kartoték ani jednotná metodika, ani soubor doporučení. Tento stav je možná zapříčiněn i způsobem práce knihovníků, kteří před správou a inovací starých lístkových kartoték mnohdy upřednostňují záznamy o nových přírůstcích.

Existence digitalizovaných lístkových kartoték přístupných online se tak stává důležitou především pro veřejnost.

Kapitola 2

Digitalizované kartotéky online, existující řešení

Definujme nejprve pojmy „digitalizovaná kartotéka“ a „digitalizace kartoték“. První termín popisuje produkt vzniklý převedením papírového média do počítačem zobrazitelné podoby, druhý označuje proces, jakým bylo digitalizace dosaženo.

Samotná podoba digitalizované kartotéky může být úzce spojena s termínem databáze, a tak je i většinou chápána odbornou veřejností. Může se ale jednat i o obrazovou kopii papírové skutečnosti, uchovávanou v počítači.

Problematika digitalizace kartoték přepisem do databázové podoby není předmětem této práce.

Termíny „digitalizovaná kartotéka“ a „digitalizace kartoték“ budou v dalším textu vždy vztaženy k produktu digitalizace skenováním doplněného případně o další atributy a relace. Kartotéka převedená do databázové podoby, včetně původní obrazové informace, je v textu označována jako „plně digitalizovaná“.

2.1 Digitalizace skenováním a částečným přepisem

2.1.1 Lístkový katalog/ Card Catalogue

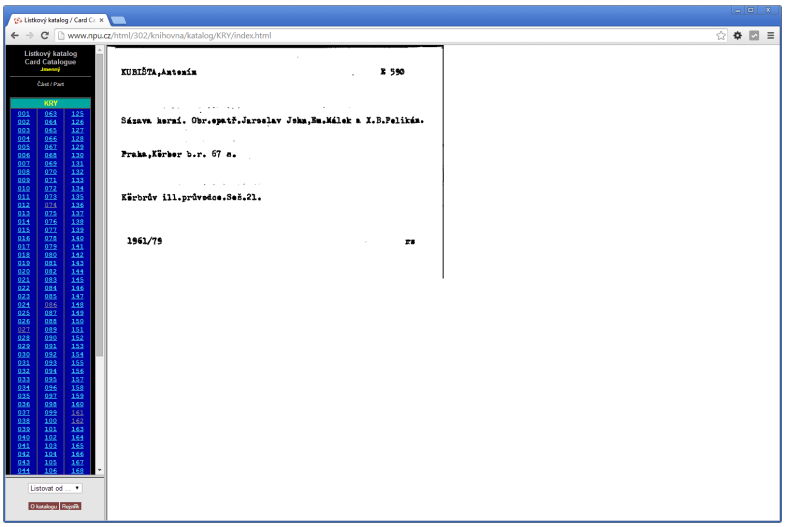
Online katalog vyvinutý Knihovnou Akademie věd kolem roku 2000 je nejjednodušším ze zde prezentovaných systémů.

Rozhraní umožňuje pomocí jednoduchých voleb zobrazit skenovanou repliku katalogů.

Ze seznamu indexů na hlavní straně volí uživatel již v dalším kroku konkrétní pořadové číslo lístku.

Autor rozhraní umožňuje zvolit uživateli jiný index kdykoli během práce s kartotékou, tím ale výčet uživatelských funkcí končí.

Kartotéka neřeší vícestránkové lístky, nenabízí ani žádný další komfort



[16]

Obrázek 2.1: Lístkový katalog/ Card catalogue, NPÚ

pro uživatele, jakým by mohlo být zvětšování či zmenšování nebo více lístků na obrazovce. V prozkoumávaných instalacích tohoto systému je k dispozici maximálně několik desítek tisíc záznamů. Zvolené rozhraní je pro takový počet postačující, ale pro větší počet záznamů by již byl pohyb v takové kartotéce pro uživatele nepohodlný.

Lístky ve zkoumaných instalacích jsou uloženy v relativně malém rozměru, do 650x450 bodů, ale i tak je jejich čitelnost vyrovnaná.

Aplikace Lístkový katalog/ Card catalogue poskytuje levný a rychlý způsob jak digitalizovat kartotéku a poskytnout ji k užívání na Internetu.

Bohužel jakýkoli další rozvoj v této podobě by byl obtížný a není s ním ani počítáno. Katalogy jsou k určitému datu uzavřeny a všechny další změny a přírůstky je již nutné hledat v nové databázové verzi[21].

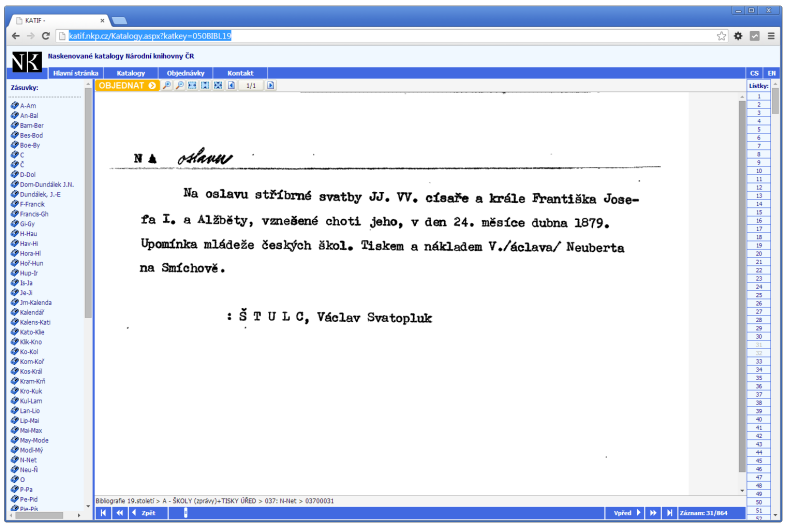
2.1.2 KATIF

Jak již bylo popsáno v kapitole 1 na straně 3, technologie RETROKON vznikla koncem devadesátých let v Národní knihovně.

Retrospektivní konverze probíhá v rámci dotačního programu VISK 5 RETROKON[30] dodnes a je do ní zapojeno na několik desítek knihoven.

Vlastní systém Národní knihovna nazvala „KATIF“ a je dostupný z webové adresy `http://katif.nkp.cz`. Je v něm nabízeno deset digitalizovaných kartoték sbírek, a to katalogy: Generální katalog I-III, Slovanská knihovna, Ruský zahraniční hist. archiv, Bibliografie 19.století, Katalog hudebních tisků, Notový incipitový katalog, Katalog knihovnické literatury, Dokumentační kartotéka knihovnické literatury, z nichž první jmenovaný je nejobsažnější.

Prvním krokem k listování kartotékou je volba šuplíku (zde pojmenován



[10]

Obrázek 2.2: KATIF, NK ČR

„Zásuvka“). Seznam šuplíků, zásuvek, je umístěn do rolovatelného okna v levé části obrazovky.

Příliš dlouhé seznamy šuplíků mohou být autorem ještě rozděleny do tak zvaných skříní. Výběr šuplíku pak probíhá na více úrovních.

U tohoto členění se již projeví nárok, jaký je kladen na uživatele při volbě kam postupovat ve vyhledávání. Uživatel se sice pohybuje po jakési pomyslné pyramidě výběru, ale díky rozlišovacím prvkům v podobě dvou a čtyřpísmenných shluků písmen toto členění klade velký důraz na jeho schopnost orientovat se v pořadí písmen v abecedě a tedy i v kartotéce.

Po rozkliknutí poslední úrovně se v centrálním okně objeví první lístek a s ním po pravé straně také číselný seznam od jedné do počtu lístků šuplíku (přibližně jeden tisíc).

Poklepem na dané číslo může uživatel šuplíkem „listovat“.

Toto ovládání je intuitivní, ale ne příliš šikovné pro opakované vyhledávání. Zručný čtenář by patrně v lístkové kartotéce byl schopen informaci vyhledat rychleji.

Z ostatních funkcí umožňuje rozhraní zvětšený nebo zmenšený pohled na lístek.

Bohužel je možné zobrazit najednou vždy pouze jeden lístek a protože záznamy není možné ukládat, uživatel je odkázán pouze na svou paměť a psané poznámky.

Velkou výhodou samozřejmě zůstává už ten důležitý fakt, že je katalog přístupný online a každý ze záznamů, kniha, je možný jedním tlačítkem objednat k výpůjčce.

Autoři kartotéky, patrně vedeni snahou uspořit místo i přenesená data,

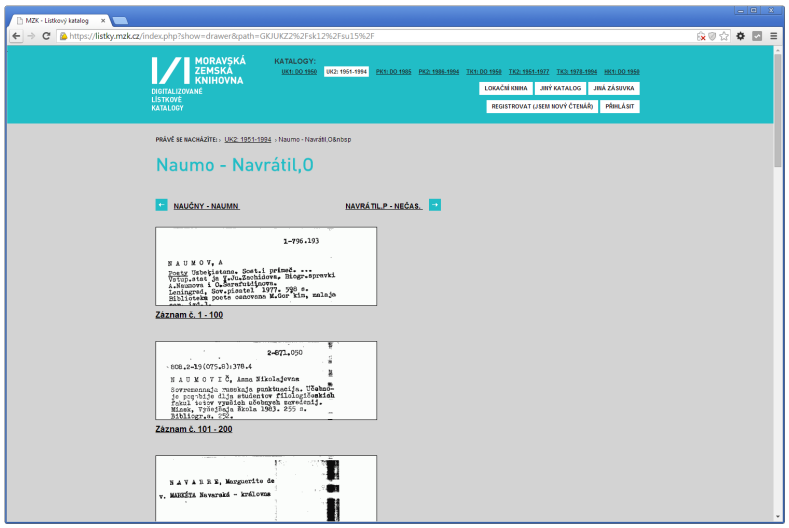
nejspíš obrázky optimalizovali a převedli do černobílého provedení. Některé zdroje, především na starším zažloutlém papíře, jsou díky tomu obtížně čitelné. Nicméně v databázi aplikace jsou lístky uloženy větší a je možné je bez újmy na kvalitě minimálně dvakrát zvětšit.

Technologie RETROKON ve své specifikaci obsahuje metodologii pro vytěžování dat z obrazového záznamu. Katalog KATIF ovšem obsahuje pouze skeny jednotlivých lístků. Další data, pokud se je z obrazových podkladů podaří získat, nejsou ukládána v katalogu KATIF, ale jsou předávána databázi ALEPH 500 (viz sekce 2.2.1 od strany 10).

Systém KATIF je podporován dotačním programem a přihlášené instituce mohou své skenované katalogy nejen zpřístupňovat touto aplikací, ale také získávat prostředky na jejich další digitalizaci.

Z uživatelů uveďme například některé knihovny Akademie věd, Fakultu sociálních věd Univerzity Karlovy v Praze, Slezské zemské muzeum nebo Jihočeskou vědeckou knihovnu v Českých Budějovicích.

2.1.3 Moravská zemská knihovna



[15]

Obrázek 2.3: Generální katalog, Moravská zemská knihovna

Velmi zajímavé rozhraní je navrženo pro osm knihovních katalogů Moravské zemské knihovny, ve kterých jsou uloženy oskenované, původně papírové, záznamy kartoték do roku 1994.

Stránky jsou přehledné, s citem pro barvu. Autoři umístili do hlavičky stránky vodorovně abecední členění a pak v širokém sloupci nabízejí rozsahy jednotlivých šuplíků.

Po rozkliknutí se zobrazí vždy částečný náhled na každý stý lístek a postupným rozklikáváním se uživatel dostává na každý desátý a pak na každý

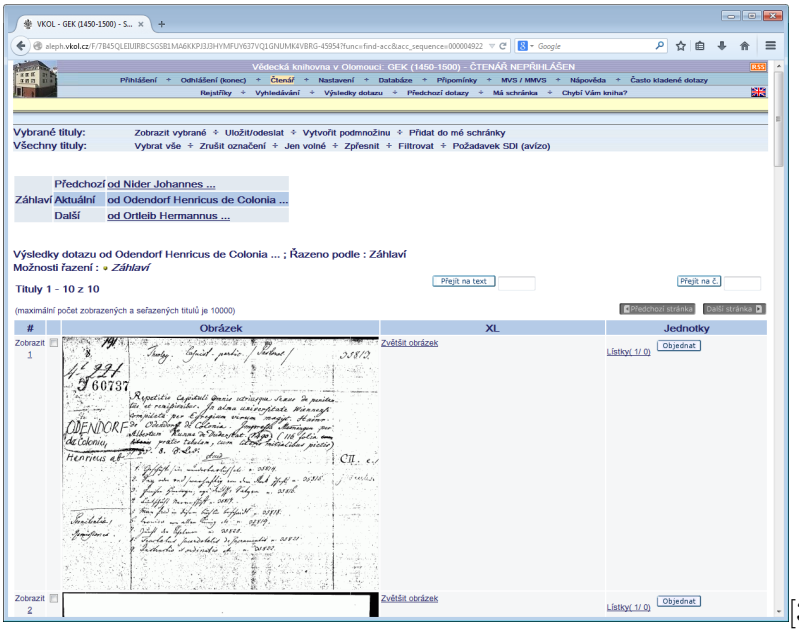
jednotlivý lístek kartotéky.

Během skoků jsou zobrazovány jen horní třetiny lístků, což usnadňuje hledání pohledem na titulek. V poslední úrovni je už lístek zobrazen celý.

I když je i zde nutnost rolovat pro výběr dalšího skoku, listování touto kartotékou je v porovnání s předchozími pohodlné a přehledné.

Dojem z celé kartotéky trochu kazí fakt, že lístek na nejnižší úrovni nejde zvětšit. Je uložen v relativně malém rozlišení, takže ani zvětšení mimo webové okno v jiném programu čitelnosti nepomůže.

2.1.4 Aleph 500



[31]

Obrázek 2.4: GEK, Vědecká knihovna v Olomouci

Jediný, v tomto výčtu zastoupený, komerční produkt umožňující procházení skenovaných lístkových kartoték je izraelský produkt Aleph 500.

Vědecká knihovna v Olomouci tímto způsobem zpracovala šest svých lístkových katalogů.

Pohyb po abecedně seřazených indexech není příliš komfortní. Probíhá vždy po skupině deseti indexů čítající deset titulů a umožňuje skočit vždy jen o deset indexů vpřed nebo vzad.

Systém umožňuje hledání podle indexů, zvětšení cílového obrázku a případně objednávku výpůjčky. Katalog je postupně přepisován do databázového formátu, ale záznamy MARC již nejsou zpětně odkazovány na původní obrazovou podobu.

Systém Aleph 500 představuje velmi komplexní a nákladné databázové řešení, další údaje jsou zmíněny v sekci 2.2.1 na straně 10.

2.2 Databázové knižní katalogy

Na českém trhu je v této chvíli k dispozici několik komerčních produktů zaměřených na digitalizaci kartoték knižních katalogů. Tyto produkty jsou čistě databázové povahy a neumožňují plynulý ani fázový přechod od lístkové kartotéky do plně digitalizované podoby.

Tato sekce může sloužit jako reference k existujícím řešení, ale protože plně digitalizované kartotéky nejsou předmětem této práce, jsou zde uváděna jen základní fakta.

2.2.1 Aleph 500

Katalog knihovního systému izraelské společnosti ExLibris představuje patrně komerčně nejúspěšnější řešení evidence knižních katalogů v České republice.

Jediným oficiálním distributorem na našem území je Univerzita Karlova. Ta také zajišťuje a koordinuje všechna nasazení a řeší s dodavatelem případné problémy a požadavky na softwarové úpravy.

V současné době je s výjimkou UK používán většinou knihoven ústavů Akademie věd, Masarykovou univerzitou, Vědeckou knihovnou v Olomouci, krajskými knihovnami v Ostravě a Hradci Králové, Moravskou zemskou knihovnou a několika dalšími institucemi.

Je to řešení ucelené a ověřené. Poskytuje řadu návazných modulů pro správu dat, uživatelů, výpůjční protokol a mnoho dalších funkcí, které zkušený knihovník ocení.

Nedocenitelnou výhodou je možnost zapojení více knihovních katalogů a sdílení autoritních bází, což usnadní práci nejen správcům, ale hlavně uživatelům při hledání.

Překážkou pro pořízení může být licenční politika, která je založena na pravidelných platbách za užívání systému podle počtu spravovaných bází a počtem souběžných přístupů.

Tento přístup je nevhodný zvláště pro malé organizace a dále pro organizace s nejasným rozpočtem budoucích období.

Překážkou mohou být i zvýšené nároky na technické vybavení a obtížnější administrace celého systému.

2.2.2 Rapid Library/ ARL

Na domácím česko-slovenském poli se společnosti Cosmotron Bohemia, s.r.o. podařilo vyvinout úspěšný knihovnický software Rapid Library, v poslední době doplněný novým knihovnickým software nazvaným ARL, Advanced Rapid Library.

Jedná se o velmi komplexní systém správy knihovních, muzejních a galerijních sbírek, postavený nad aplikačně-databázovým systémem Caché.

Stejně jako předchozí produkt, obsahuje i ARL řadu modulů vhodných pro oběh dokumentů, evidenci čtenářů, statistiky, slovníky, fakturace, automatizaci výpůjček, vše postavené na OPAC/ IPAC pracujících s katalogem.

O tom, že se jedná o velmi důležitého a solidního domácího dodavatele této technologie snad nejlépe svědčí fakt, že jeho zákazníky jsou Národní knihovna, Středočeská vědecká knihovna Kladno, Památník národního písemnictví, Parlamentní knihovna, Nejvyšší soud České republiky Brno a řada dalších.

2.2.3 KpWinSQL

Tento produkt patří bezesporu k nejvýznamnějším v oblasti katalogizace. Je vhodný pro všechny typy knihoven – veřejné, odborné, univerzitní, školní, archivy a muzea[13].

Společnost KP-SYS spol. s r.o. vytvořila modulárně bohatý a cenově přístupný produkt používající základní knihovnický formát MARC 21 umožňující běh SQL serveru na platformách MS Windows, Linux, Unix, Mac OS X, FreeBSD, Solaris.

V této chvíli systém KpWinSQL využívá celkem 288 knihoven a dalších 252 knihoven zapojených do regionálních systémů[13].

K nejdůležitějším zákazníkům patří Národní muzeum, Divadelní ústav, ale i organizace jako jsou České dráhy nebo Český Hydrometeorologický ústav.

2.2.4 Tritius/ Clavius/ LANius

Tři jména v nadpisu reprezentují retrospektivně tři systémy vyvinuté společností LANius s.r.o.

Chronologicky nejstarší systém LANius je navržen pro prostředí DOS a vyvinut byl nástrojem MS FoxPro 2.5a. Výrobce uvádí, že tento systém používalo, a je možné, že z části ještě používá, 109 institucí, především místních a obecních knihoven a muzeí.

První z pokročilejší verze systému, nazvaná Clavius, byla navržena pro prostředí Windows již v roce 1997 a k jejímu vývoji byl použit nástroj MS Visual FoxPro.

Od doby prvního nasazení dosáhl počet instalací úctyhodného čísla 2939[12] a datum poslední revize z prosince tohoto roku dává tušit, že je o systém stále zájem.

Základ systému se postupem doby změnil na OPAC Carmen a v poslední době bylo rozhraní Clavius nahrazeno rozhraním Tritius[25].

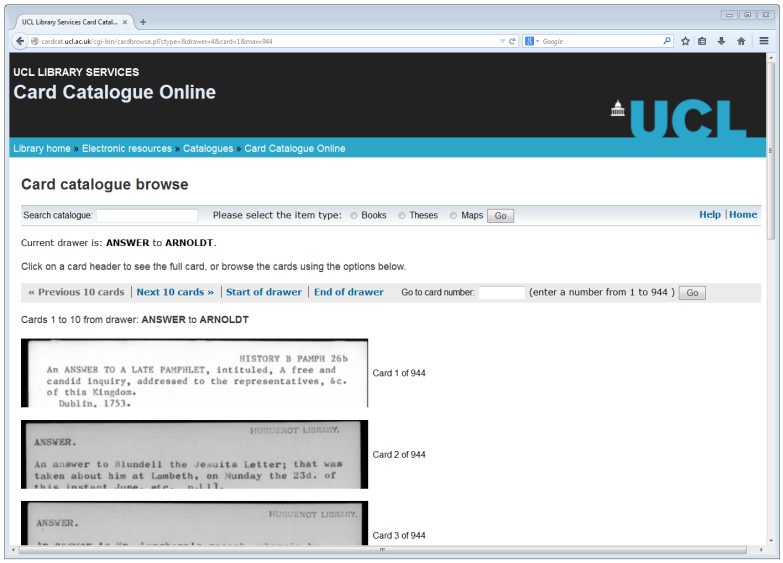
2.3 Situace v zahraničí

Při zjišťování zahraničních zdrojů jsem se soustředil na hledání skenovaných knižních katalogů na webových stránkách britských, amerických, německých a rakouských univerzitních a vědeckých institucí. Prezentace často obsahují heslo „card catalog“ a uvádějí i postup, jak se v informacích kartotéčního lístku orientovat, ale pro hledání odkazují zájemce na fyzickou kartotéku, případně uvádějí odkaz na databázový přepis.

Otázku převodu lístkových kartoték do digitální podoby začaly řešit zahraniční instituce oproti nám dříve, v době ekonomického růstu, a tedy za jiných finančních podmínek. V kontextu k politické situaci je třeba uvést rok 1989 jako mezník, od kterého se podstatně zjednodušil přístup našich institucí k výpočetní technice.

Skenované kartotéky nejsou v zahraničí tak časté, pravděpodobně díky brzkému přepisu dat z lístkových katalogů pro znakově orientovaná zařízení.

2.3.1 Card Catalogue Online/ Velká Británie



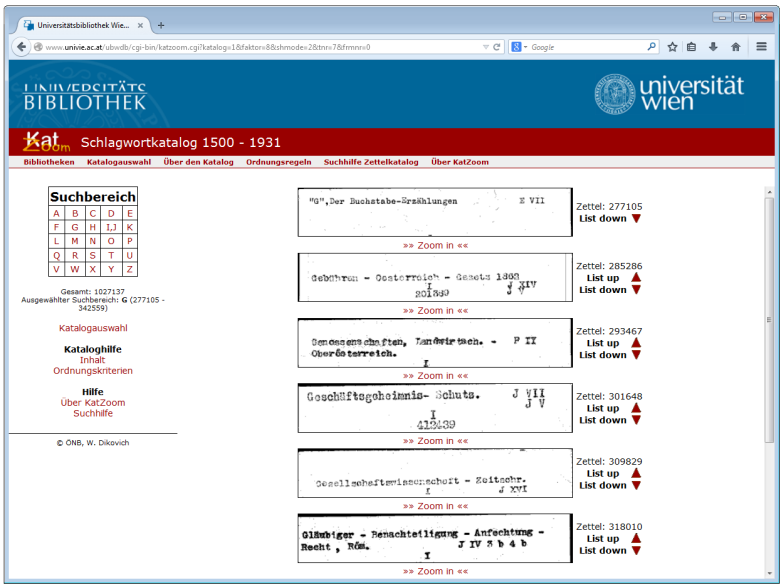
[26]

Obrázek 2.5: UCL Library Services, University College London

Skenovaný katalog University College London obsahuje záznamy do roku 1980[27], od tohoto data jsou vedena databázově.

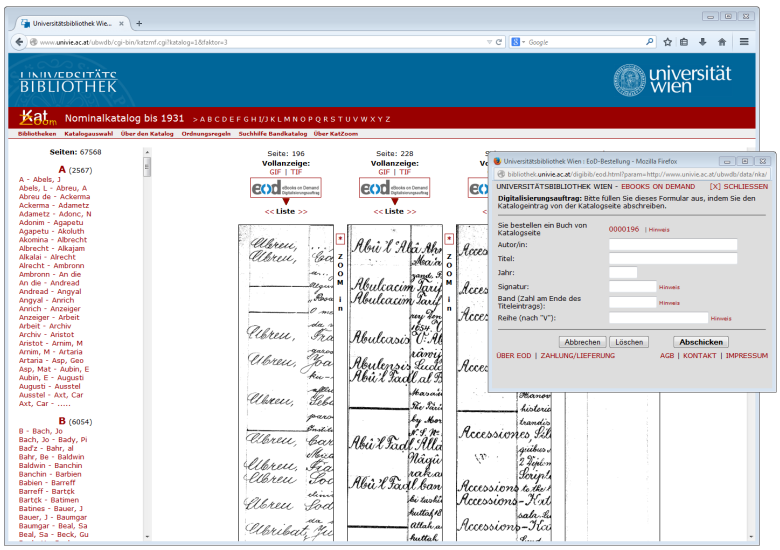
Uživatel může zvolit ze tří typů záznamů. Books, Theses a Maps. Dále je k dispozici vyhledávací okno, které po zadání hledaného hesla přiblíží nejbližší písmenný index k výběru záznamu. Indexy jsou rozděleny do oddílů po přibližně 1000 záznamech, tlačítka nabízejí posun o deset záznamů, skok na začátek a konec indexu a také zadání přesného čísla v rámci indexu.

2.3.2 KatZoom/ Rakousko



[29]

Obrázek 2.6: KatZoom, volba písmene, Universität Wien



[28]

Obrázek 2.7: KatZoom, volba indexu, Universität Wien

Rakouskou národní knihovnou (Österreichische Nationalbibliothek) byl vyvinut software KatZoom. Systém převzalo několik knihoven a v současné

době je systémem spravováno celkově 12 000 000 kartotéčních lístků, největší báze má velikost 1 690 000 záznamů[11].

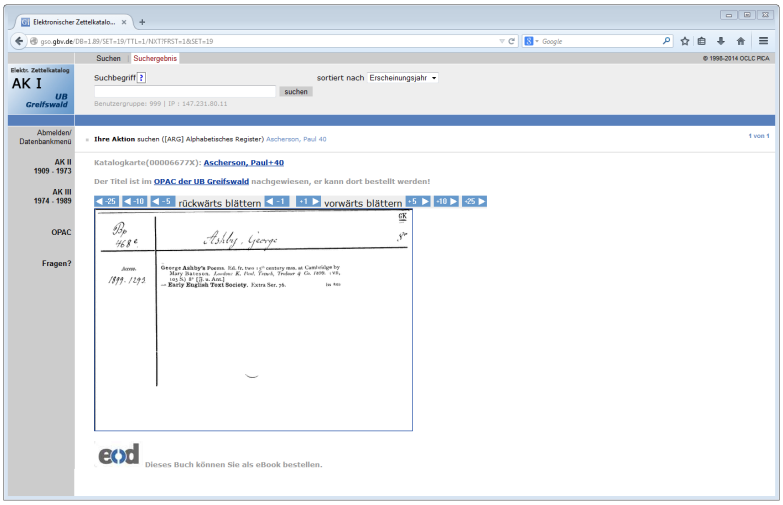
Pokud knihovna disponuje více katalogy, systém umožňuje výběr jednoho z nich. Následuje výběr indexu podle četnosti záznamů buď volbou písmene (obrázek 2.6) nebo volbou indexu písmene (obrázek 2.7).

Uživateli se zobrazí vrchní část lístku s krokem vypočítaným tak, aby se na obrazovku vešlo vždy devět záznamů. Volbou „Zoom in“ volí uživatel další vypočítanou úroveň. Případně jsou k dispozici volby „List up“ a „List down“ pro skok na sousední skupinu záznamů. V nejnižší vrstvě jsou zobrazeny lístky celé, bohužel bez možnosti zvětšení.

V některých katalozích jsem našel funkci uživatelského přepisu základních bibliografických údajů do strukturovaného záznamu s možností odeslání, jak je vidět na obrázku 2.7.

Celý systém působí velmi přehledně, i když občas rušivě zapůsobí náhlá změna layoutu stránky.

2.3.3 Elektronischer Zettelkatalog AK I-III/ Německo



[8]

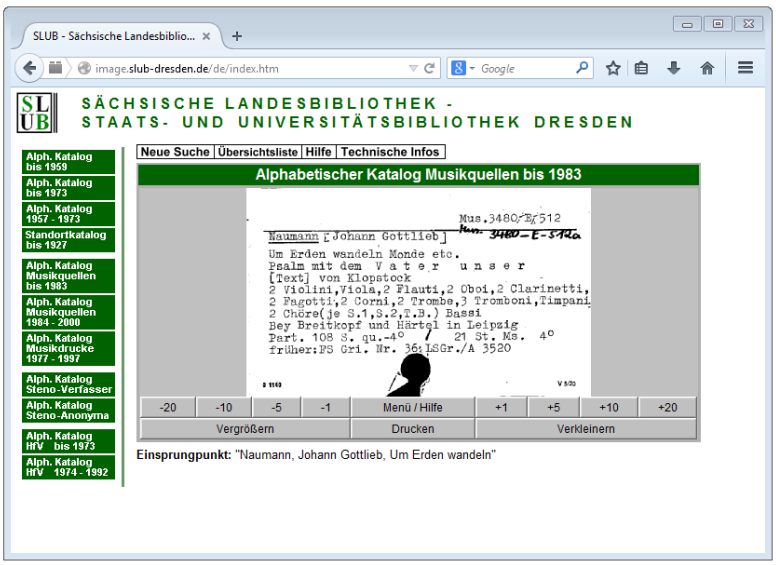
Obrázek 2.8: AK I, Universitätsbibliothek Greifswald

Katalog Greifswaldské univerzitní knihovny je rozdělen na tři části podle období. Uživatel může v první fázi rozhodnout, zda bude katalog řazen podle roku vydání nebo podle důležitosti. Dále je do vyhledávacího okna zadáno jméno autora a systém nabídne odkaz do indexu autorů, kteří jsou rozlišeni pouze číslem.

V následujícím kroku je zobrazen jeden celý lístek a tlačítka posunu o 1, 5, 10 a 25 lístků vpřed nebo vzad. Lístek je možné zvětšit otevřením přes celé okno prohlížeče. Bohužel je pak nutné vrátit se tlačítkem Zpět.

Na stránce jsou i odkazy do databázového katalogu OPAC greifswaldské univerzity a v první části katalogu i ikona pro objednání elektronické verze knihy.

2.3.4 Digitalisierte Zettelkataloge der Sächsischen Landesbibliothek/ Německo



[23]

Obrázek 2.9: Digitalisierte Zettelkataloge der Sächsischen Landesbibliothek, Staats- und Universitätsbiliothek Dresden

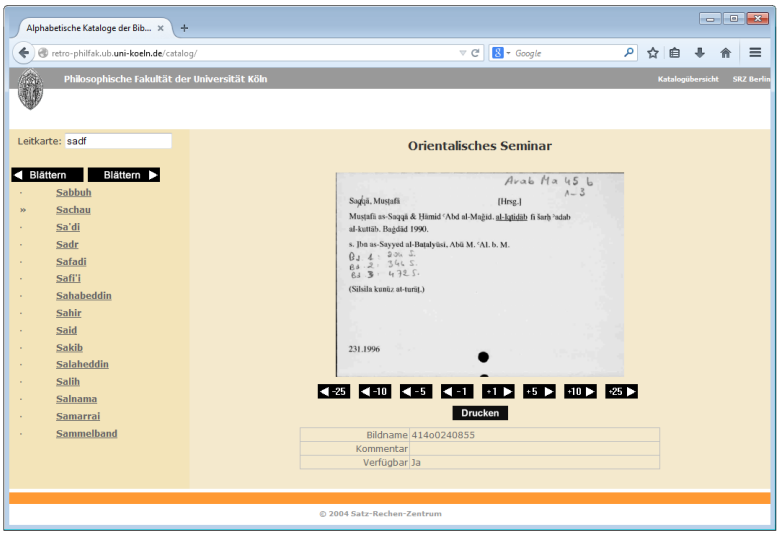
Většina fondů Státní a univerzitní knihovny v Drážďanech je již převedena do online OPAC katalogu, nicméně knihovna stále provozuje systém skenovaných katalogů ve stavu, v jakém ho nabídla uživatelům v roce 2000[23].

Systém nabízí k výběru v levém sloupci jeden z jedenácti katalogů. V dalším kroku uživatel zadá jméno a zobrazí se seznam s vyznačeným indexem, zároveň také šest indexů předchozích a šest následujících indexů. Poklepnutí na index vyvolá obrázek prvního lístku. Ten se zobrazí přes šířku zobrazovacího okna, které je ale v závislosti na rozlišení monitoru výrazně menší než okno prohlížeče. Pokud je lístek delší než zobrazovací okno, je možné ho zvětšit, ale uživatel na informace mimo zobrazovací rámeček není upozorněn.

S obrázkem se zobrazí i nabídka tlačítek posunu o 1, 5, 10 nebo 20 lístků vpřed a vzad. Nechybí tlačítka pro zvětšení a zmenšení, také je zde umístěno tlačítko pro tisk a menu. Další nabídka obsahuje funkce otočení, změnu kontrastu, zobrazení celého lístku a některé další.

Katalog pro správnou funkci potřebuje Java-applety a aplikace je citlivá i na druh Internetové prohlížeče.

2.3.5 Alphabetischer Katalog, Universität Köln/ Německo



[2]

Obrázek 2.10: Alphabetischer Katalog, Philosophische Fakultät der Universität Köln

Všechny katalogy Univerzity v Kolíně nad Rýnem jsou nabídnuty v jednom sloupci uprostřed okna prohlížeče.

Po výběru se uživateli změní okno. V levém sloupci je okno vyhledávání pro zadání indexu a seznam indexů od písmene A nebo od vyhledávaného jména. Prostřední část okna prohlížeče vyplňuje lístek a ovládací tlačítka pro přesun o 1, 5, 10 a 25 lístků vpřed a vzad, tlačítko tisku a také informace o jméně souboru obrázku, komentář a dostupnost knihy v knihovně.

Lístek je možné poklepáním zvětšit na maximální velikost.

I přes velmi málo funkcí je pohyb po kartotéce rychlý, jednoduchý a intuiivní.

Kapitola 3

Obecný postup převodu lístkových kartoték

Schematický postup zachycuje diagram na obrázku 3.1.

3.1 Analyzování cílů, možností a prostředků

Přecházíme-li od lístkových kartoték ke kartotékám digitálním, měli bychom si uvědomit, že se změnou média měníme i náš náhled na data. Pro většinu uživatelů je zprvu nepochopitelná představa jiného užitku, než jaký získávali z původní lískové kartotéky. Volí řešení založená na svých poznatcích, vyžadují skeny původních lístků a nástroje procházení ekvivalentní těm v reálném světě.

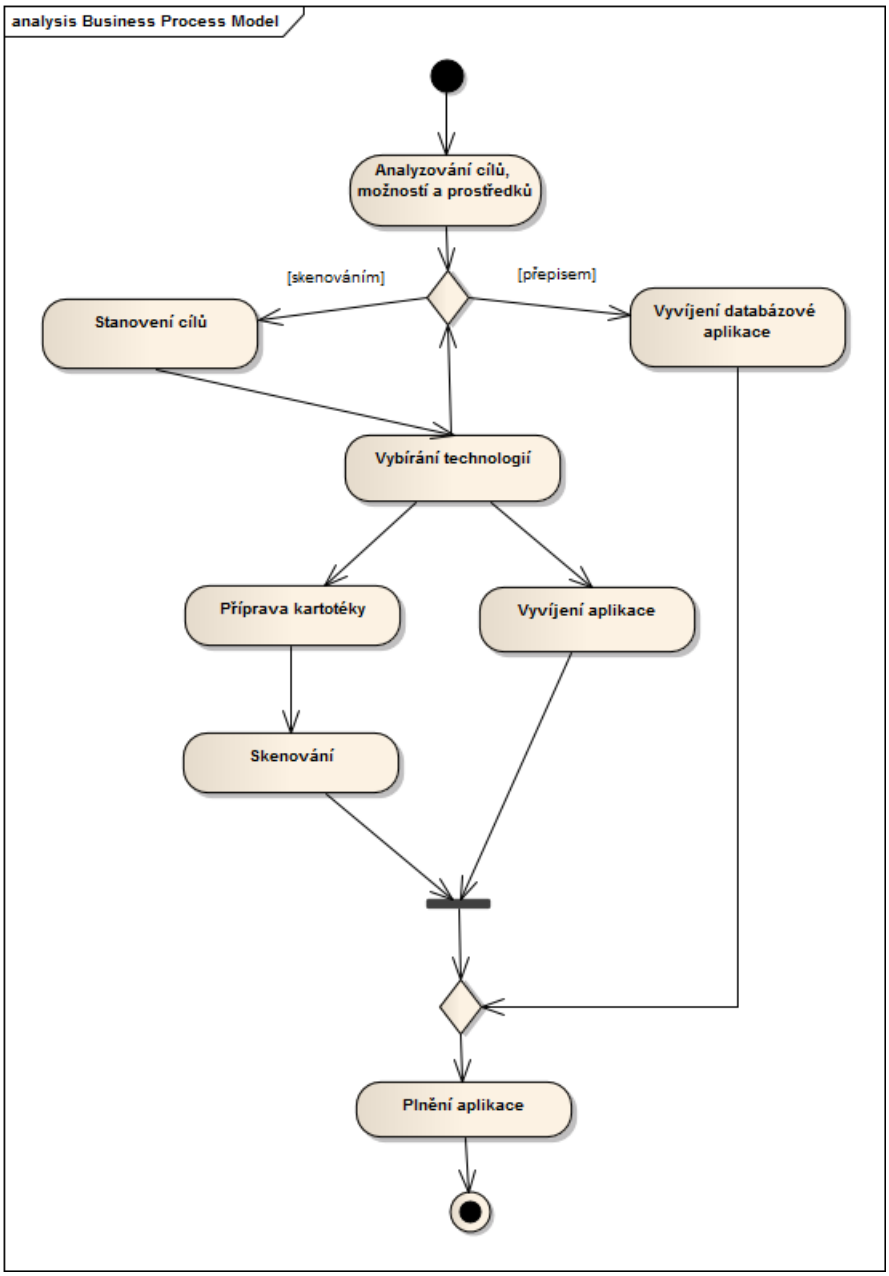
Za předpokladu, že výpočetní technice poskytneme data v příslušném formátu, získáme v krátkém čase celou řadu zajímavých výstupů. Lístkové kartotéky taková data obsahují a každý, kdo je s možnostmi databází seznámen, začne okamžitě uvažovat o přepisu do databázové podoby.

Pochopením možností často vzniká opačný extrém, a to snaha o plnou digitalizaci kartotéky s postihnutím všech relací, i když objem nutné práce vůbec nemusí odpovídat výslednému efektu a jev, který je takto podchycován je v kartotéce zastoupen minoritně nebo na něj není nikdy vznesen dotaz.

Přepis do digitální podoby se tak může stát náročnějším úkolem, než bylo fyzické vytvoření kartotéky původní.

V této fázi přípravy převodu by měl být projekt posouzen ze tří propojených pohledů. Odborným náhledem na obsah kartotéky, na potřeby uživatelů a na možná rozšíření, která výpočetní technika přinese. Pohledem programátora posuzujícího možnosti, časovou i technickou složitost. Dohromady pak posouzením finanční náročnosti.

Strany by měly být schopné kvalifikovaně odhadnout objem a časovou náročnost práce, která bude potřeba na jejich straně problému vykonat. I haléřové položky mohou při statisícových objemech dat vážně narušit rozpočet.



Obrázek 3.1: Model převodu lístkových kartoték

Je potřeba učinit zásadní rozhodnutí, zda kartotéka bude přepsána do databáze nebo bude oskenována a jen určitá pole se stanou atributy budoucího záznamu.

3.2 Stanovení cílů

Prostor dalších úvah můžeme rozložit mezi dvě krajní polohy. Jednou je oskenování bez dalšího zásahu a druhou převod do databázové podoby.

Oskenování bez dalšího zásahu

Fyzická kartotéka je intuitivně chápána množinově jako systém podmnožin

$$\textit{lístek} \subset \textit{šuplík} \subset \textit{kartotéka}.$$

Tento pohled není zcela přesný. Nutí uživatele chápat záznamy jako osamocené ostrovy informací, které spolu zdánlivě nemají nic společného.

Skutečnost ale spíše odpovídá představě, kdy kartotéku tvoří jeden velmi dlouhý šuplík naplněný lístky, které mají čas od času mezi sebou zarážku pro snazší vyhledávání.

Na základě tohoto pohledu by bylo možné lístky jednoduše oskenovat a vhodným průběžným očíslováním nebo propojením vytvořit digitální kartotéku, která pro dosažení funkčnosti fyzické nebude potřebovat více než jednoduchý prohlížeč, případně pomocnou nástavbu indexů označující začátky původních šuplíků.

Klady Digitalizace kartotéky tímto způsobem je velmi efektivní na čas, práci i finanční prostředky a kartotéka může být v krátké době přístupná všem zájemcům na Internetu. Nedojde k žádné ztrátě informace.

Zápory Uživatelé získají navíc pouze přístup k informacím přes Internet nebo jiné elektronické médium. Doplňování a opravy záznamů jsou obtížné a časově výrazně náročnější než v původní papírové kartotéce.

Převod do databázové podoby

Záznamy fyzické kartotéky podrobíme detailnímu zkoumání a analýze.

Na základě databázových metod a postupů stanovíme atributy, vyznačíme relace, naprogramujeme databázi a záznamy kartotéky do ní lístek po lístku přepíšeme.

Klady Vyhledávání je otázkou zlomku času oproti vyhledávání v obrazové formě.

Je možné získat data tříděním podle libovolného klíče z libovolné podmnožiny dané báze, na základě relací vyhledávat souvztažnosti.

Veškerá data jsou plně propojena a validována. Existuje číselník validovaných atributů.

Je velmi snadné záznamy opravovat, editovat i přidávat nové.

V porovnání k obrazovým datům jsou textová data výrazně menší.

Zápory Proces je nesmírně náročný na čas, práci i finanční prostředky.

Samotný proces vytvoření databáze předpokládá úzkou spolupráci a pochopení jak na straně zadavatele, tak na straně programátora.

Převod klade velké nároky na odbornou způsobilost operátorů, kteří data přepisují. I přes velké úsilí a křížové kontroly dochází k chybám, které vzhledem k zániku původního záznamu již nikdo není schopen dohledat a odstranit. Čas potřebný k přepisu i řádově malé kartotéky je obrovský.

Provedená práce nemusí odpovídat výsledné užitné hodnotě databáze. Pracně budované relace nemusí být nikdy využity.

Dojde ke ztrátě nekatalogizovatelných a netypických atributů, jako jsou poznámky uživatelů kartotéky na okrajích karet, barva inkoustu značící určité období i způsob zpracování dat a jiné abnormality záznamu.

V reálném světě navíc knihovník raději přepisuje údaje do databáze podle knihy než podle hotového kartotéčního lístku.

Srovnání

Obě varianty mají velmi přesvědčivé kladné a naopak velmi odrazují záporné vlastnosti.

Holé oskenování kartotéky a jednoduchý nástroj pro její procházení určitě splní základní funkce tak, jak byly naplněny u fyzické kartotéky. S možnostmi, které výpočetní technika nabízí to ale není mnoho. Za nízkou cenu je mnohdy možné kartotéku doplnit o další funkčnost a zvýšit tak komfort vyhledávání i výtěžnost informací.

Je třeba stanovit, jaké nároky budeme na budoucí kartotéku klást, kdo jsou a budou její uživatelé a jaké otázky jejím prostřednictvím budeme chtít řešit.

Zadavatel si většinou zprvu nedokáže představit jaké dotazy může kartotéce klást. Je zvyklý na papírovou verzi a práci s ní. Proto se bude ze začátku klonit k variantě prostého skenu bez dalšího zásahu. Jakmile se začnou probírat úkoly, které s kartotékou byly v historii řešeny, vynoří se řada zajímavých dotazů, které jsou zpracovatelné pouze v plně databázové podobě a zadavatel začne uvažovat o opačném extrému, na který ale dost často nejsou prostředky ani čas.

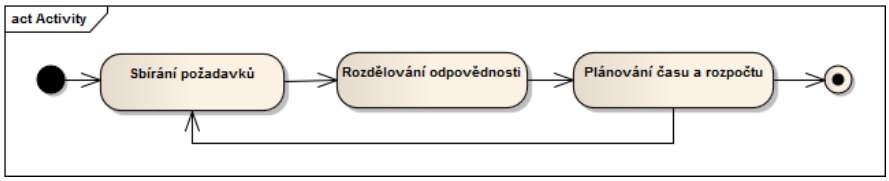
Požadavky je třeba určitou dobu shromažďovat a třídit podle důležitosti, která se bude u většiny kartoték měřit četností dotazu daného typu požadavku.

Triviálním příkladem může být vyhledávání podle jména. Naplnit atribut „jméno“ pro každý záznam může být už u několika set tisíc záznamů práce na mnoho měsíců. Přitom se záznamy opakují a operátor by zbytečně opakovaně zadával stejné jméno. Cíl může být stanoven na rozdělení kartotétky po

několika stech záznamech. Popis se řádově zjednoduší a v objemu několika set lístků uživatel najde hledaný záznam také poměrně snadno.

Rozumnou investicí tak můžeme dosáhnout lepšího výsledku a uživatelům kartotéky nabídnout více služeb, než by jim poskytla pouze naskenovaná kartotéka.

Střízlivým rozborem požadavků tak docílíme lepších výsledků než jaký poskytne prostý sken a ekonomické úspory proti převodu do databázové podoby, která může být v určitém směru plýtváním času i peněz.



Obrázek 3.2: Model stanovení cílů

Chyby ve stanovení cílů mohou vést k nenapravitelným škodám při vývoji aplikace a mohou se zřetelně podepsat na výsledku. Je nezbytné, aby v této fázi úzce spolupracovali jak odborníci, kteří kartotéku znají a jsou detailně obeznámeni s problematikou celého oboru, tak odborníci na výpočetní techniku, kteří mají představu o náročnosti kódování jednotlivých cílů a o jejich faktické splnitelnosti.

Je třeba postupovat tak, jak je naznačeno na obrázku 3.4. Spolu s cíli rozdělit odpovědnost za jejich plnění, odhadnout časový plán a finanční náročnost. V případě nemožnosti naplnění je nutné celý proces zopakovat a cíle pozměnit, případně zajistit více času nebo peněz.

3.3 Vybírání technologií

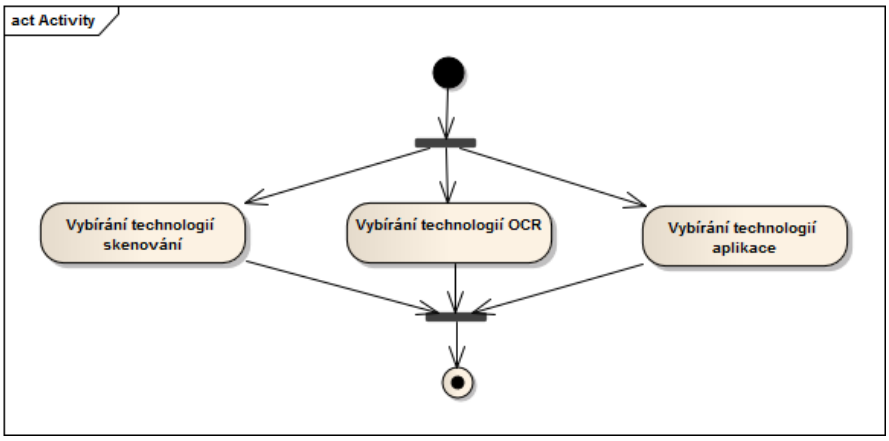
Vybírání technologií skenování Podle možností zadavatele může převod do obrazového tvaru proběhnout na různém zařízení.

Pravděpodobně nejlepší volbou pro skenování lístkových kartoték je jednoprůchodový skener.

Rozsáhlejší kartotéka vyvolá potřebu rychlejšícho skeneru. Je třeba přihlédnout k velikosti i ke stavu digitalizovaného média. Pokud je kartotéka tvořena lístky stejného druhu, stačí aby skener dokázal zpracovávat jeden druh papíru.

Kartotéka může být naplněna záznamy tvořené lístky různé kvality, od pauzovacího papíru přes papír recyklovaný, kartonové kartičky 200g/m^2, makulaturu s potrhanými rohy a samozřejmě dnešní kancelářský papír gramáže 80g/m^2 nevyjímaje.

Je proto třeba velmi pečlivě volit podavač, v ideálním případě si přístroj před pořízením vyzkoušet.



Obrázek 3.3: Model vybírání technologií

Lístky jsou často popisovány v kartotékách oboustranně, a proto investice do jednoprůchodového oboustranného skeneru nebude vzhledem k ceně pracovní síly vyhozenými penězi.

Rychlost skenování při určitém rozlišení je určujícím faktorem pro optimální volbu skeneru. Pokud skenované zdroje nebudou podrobeny OCR vyhodnocení, není důvod lístky skenovat ve výrazně větším rozlišení, než jaké bude následně publikováno na Internetu.

Pokud ale tento krok vynechán nebude, je více než vhodné spočítat si čas, který daný skener potřebuje pro oboustranné skenování v daném rozlišení.

Test u prodejce na signifikantím objemu dat může potvrdit vhodnost i nevhodnost výběru.

Zdržení může přivodit i technologie přenosu dat do počítače, kdy skener sice data pořídí poměrně rychle, ale počítač je není schopen přes (např. USB 2.0) stejně optimálně ukládat. U skeneru je vhodné zaměřit se na údaje o doporučeném objemu kopií za jednotku času a na cenu a dostupnost náhradních dílů, zvláště pro podavač. Doba nutná k dodání náhradního dílu nebo zařízení může celý proces výrazně zpomalit.

Volba jiného zařízení, jakým je deskový skener nebo fotoaparát, není pro digitalizaci lístkových kartoték vhodná.

Čas potřebný k manipulaci s lístkem, jeho sesnímání a zařazení se, vzhledem k rozsahu dat na lístcích obsažených, blíží času holého přepisu záznamu.

U řídkých záznamů ho i dalekosáhle přesáhne.

Volba této technologie by byla časově i finančně neekonomická.

Vybírání technologií OCR Pokud nebude zvolena krajní varianta „oskenování bez dalšího zásahu“, pro jednoduché vytěžení nestrukturovaných dat z lístku bude vhodné vybrat program OCR. Programy tohoto druhu nebývají levné a pečlivá příprava a testy před dalšími kroky se rozhodně vyplatí.

Je také možné zvolit cestu externího dodavatele skenů a externího dodavate OCR. I v tomto případě se vyplatí provést výpočty a požádat dodavatele o zkušební spolupráci nad přípravou signifikantního vzorku dat. Na první pohled neviditelné nevýhody se mohou projevit už při testech spolupráce. Dohled nad pečlivostí a kompletností provedených prací je výrazně menší a riziko poškození fyzických záznamů výrazně větší. Příkladem takového skrytého problému může být požadavek externího dodavatele na polepení každého lístku kartotéky čárovým kódem.

Vybírání technologií aplikace S ohledem na stanovené cíle je třeba zvolit detaily naprogramování aplikace a technologii, na kterém bude výsledek uložen. Především se jedná o rozhodnutí, jak bude vypadat model aplikace (klient-server, jednouživatelský apod.), volba databáze a programovacího jazyka. V závislosti na těchto volbách je třeba navrhnout hardwarové řešení s dostatečnou silou a kapacitou.

Je možné, že bude opět nutné přehodnotit původní záměry a vrátit se zpět ke stanovení dosažitelných cílů.

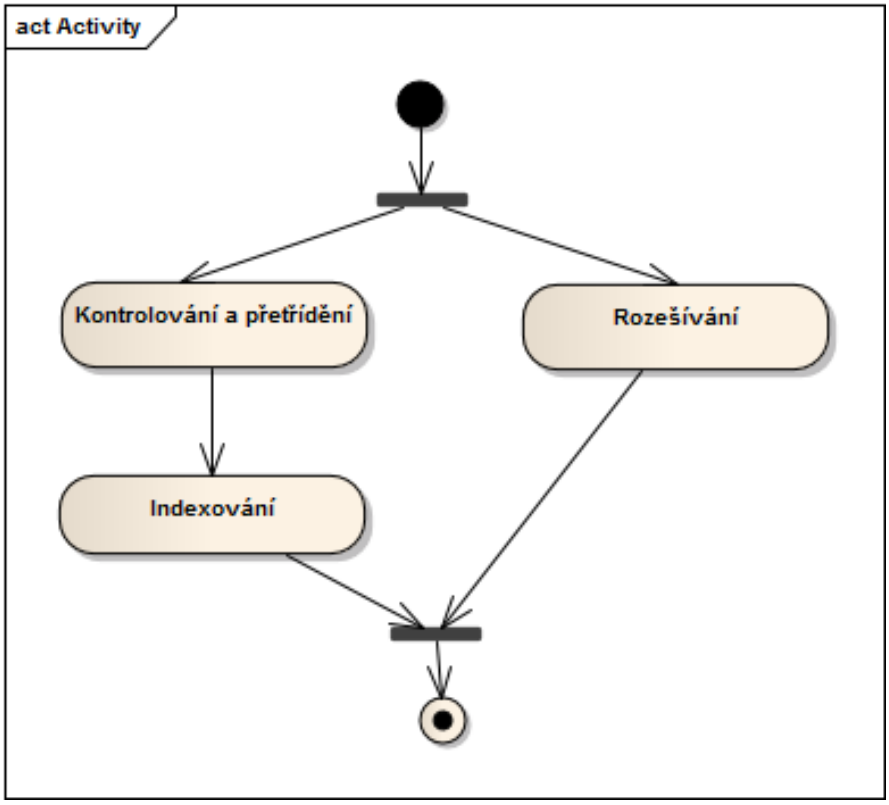
3.4 Příprava kartotéky

Proces skenování je psychicky náročný a musí být proto pokud možno co nejjednodušší bez nutnosti provádět průběžné kontroly a jiné činnosti. Je proto vhodné před začátkem celou kartotéku projít a na skenování ji připravit. Pokud je to možné, měla by být provedena základní kontrola zda jsou lístky správně orientované, nejsou zpřeházené nebo otočené.

V případě, že záznamy jsou tvořeny více jak jedním lístkem, jsou zpravidla sešity kovovou sponou nebo dočasně spojeny kancelářskou sponkou. Setkal jsem se i s kartotékou, kde záznam tvořily lístky slepené či sešité ve hřbetu do podoby sešitku. Tyto shluky je třeba před skenováním rozdělit odstraněním sponky, odříznutím kraje, případně jiným způsobem.

Pro účely skenování je vhodné v šuplíku fyzicky vyznačit pozici vícestránkových lístků pro další zpracování. V praxi se osvědčilo takové lístky pootočit o 90 °, v případě opakujících se vícelístkových záznamů mezi ně vložit zarážku v podobě barevného proužku papíru. Označení je možné i jiným vhodným způsobem, například vložením barevného proužku jiné barvy.

Rozvržení indexů Abecední členění indexů je obecně platné pro širokou paletu různých kartoték. Jemnější dělení je třeba vždy předem zhodnotit a na pokusné množině dat vyzkoušet. Indexovat lístky po několika desítkách je sice z pohledu dalšího využití při hledání záznamu výhodné, ale bude-li nastaveno příliš jemně, nemusí uživateli přinést o tolik větší komfort při vyhledávání, ale výrazně zpomalí operátora při fyzickém skenování lístků.



Obrázek 3.4: Model přípravy kartotéky

Záznamy je třeba rozdělit v rámci šuplíku na rozumně velké sekce čítající 200 až 300 lístků. Tato hodnota byla stanovena na základě testů prováděných s uživateli jako subjektivně nejoptimálnější množství. Jedná se o rozumně velkou jednotku „adresovatelnou“ v rámci kartotéky klíčem primárního atributu, jednotku tak velkou, aby uživatel i bez dalšího třídění byl schopen v ní vyhledat v rozumném čase požadovanou informaci.

3.5 Skenování kartotéky

Časově náročná činnost skenování vyžaduje dobrou předchozí přípravu a detailní rozmyšlení každé nestandardní situace, která může v průběhu skenování nastat, v ideálním případě jim předcházet.

Pokud je vše dobře připraveno, je třeba aby operátor pouze dohlédl na správnou činnost skenovacího zařízení a obslužného počítače a aby správně označoval skupiny a vícelístkové záznamy.

Skenování je práce náročná na čas, prostor, prostředky, ale především na obslužný personál, který musí mít příslušné odborné znalosti, na které bezpodmínečně musí navazovat i pečlivost a důslednost.

Svěří-li zadavatel skenování externímu subjektu (takovéto služby poskytují například společnost Xerox nebo Hawlett Packard), může tím získat velmi kvalitně oskenovaný zdroj za fixní cenu, ale zároveň může o mnohá cenná data přijít vzhledem k nezainteresovanosti zaměstnanců externího subjektu.

Zkušenosti ukazují, že zadavatelé velmi neradi svěřují kartotéky mimo území pracoviště, byť jen na předem definované období,

Nesmírná práce vynaložená na vytvoření kartotéky a její snadná zranitelnost a nenahraditelnost toto jejich chování více než ospravedlňují.

Průběžné číslování a pojmenování

Zařízení umožňující opakované pořízení obrazového záznamu vytváří soubory většinou odlišené v názvu číselným indexem, ideálně s možností nastavení počátečního indexu.

Šuplíky skenujeme po připravených blocích jednoho indexu. Blok pojmenujeme jménem indexu a odlišíme pořadovým číslem.

Není podstatné, zda číslo indexu bude průběžné nebo pro každý index znovu inicializované. Index sám tvoří primární vyhledávací klíč a pořadové číslo v rámci indexu tvoří sekundární klíč.

Takovou strukturu můžeme vytvořit jako strom adresářů, jejichž názvy jsou tvořeny názvy šuplíků, podšuplíků a indexů. Nebo tak, že název souboru je tvořen kompletním klíčem.

Těchto způsobů je bezpočet a jejich volba záleží na zvolené strategii aplikace.

V této fázi je pouze potřeba zajistit, aby u každého záznamu bylo možno zjistit jeho předchůdce, následníka a případně skupinu, do které patří.

OCR

Pouze oskenované obrazy jinou než vizuální informaci do digitalizované kartotéky nevnášejí.

Nadějí pro automatické vytěžení dat mohou být strojem psané kartotéční lístky, které se v kartotékách objevují již od dvacátých let dvacátého století

Na takto psané lístky je možné použít technologii OCR a alespoň v dále nestrukturované podobě získat velkou část textové informace.

Pokud jsou záznamy na lístky zapisovány v určitém předem dohodnutém a dodržovaném tvaru, například odděleny řádkem nebo jiným prvkem, může být tato skutečnost využita pro další heuristiku vytěžení dat. Výsledek sice nebude strukturovaný, ale takto získané segmenty výrazně zvýší kvalitu vyhledávání.

Textový OCR soubor může být následně uložen ve stejné struktuře a pod shodným názvem, jako oskenovaný obrázek.

Jeho další zpracování zajistí aplikace.

3.6 Vyvíjení aplikace

Jak již bylo řečeno, máme-li zdigitalizovaný systém lístků připravený tak, že ke každému lístku známe jeho předchůdce i následovníka, nebude problematické vytvořit aplikaci, která by uživateli zobrazila konkrétní lístek a na přání zobrazila záznam vzdálený o daný počet lístků před nebo za konkrétní pozicí.

3.6.1 Pohyb po kartotéce

Je účelné uživateli pohyb po kartotéce maximálně usnadnit.

Indexy máme již nyní data rozdělena do stromové struktury, kdy nalezení konkrétní relativně malé skupiny nezabere víc než pár skoků ve vyhledávacím stromu.

Je třeba vizualizovat tento vyhledávací strom, kde z každé úrovně musí být možnost skočit zpět nebo si vybrat úroveň nižší.

Jakmile zvolíme cílový index, je třeba zobrazit cílové záznamy opět s rozumným krokem tak, aby dalším půlením intervalu bylo možno rychle dosáhnout cílového záznamu.

Ze zkušenosti uživatelů je známo, že k nezbytnému pohybu po jednom lístku je žádán také skok na začátek a konec indexu, dále pak definovaný skok „rozumné“ velikosti, zpravidla 1, 10, 20, 50 nebo 100.

3.6.2 Úpravy v kartotéce

Odstranění záznamu

Požadavek na smazání záznamu zahrnuje pouze nutnost odstranit lístky příslušného záznamu a případně informovat předchůdce a následníka mazaného prvku, že dochází k této změně.

Problém může nastat u prvního záznamu indexu a to tehdy, je-li tento záznam jediný ve skupině. V takovém případě musí aplikace zajistit přejmenování celého indexu.

Přesun záznamu

Ve vytvořené aplikaci je třeba dořešit otázku pohybu lístku uvnitř indexu i uvnitř kartotéky.

Je zřejmé, že pohybuje-li se lístek uvnitř indexu, je třeba vhodným způsobem lístky přečíslovat, případně je v aplikaci řetězit systémem předchůdce-následník.

Pokud je nutné lístek přeřadit mezi šuplíky, je nezbytné zajistit, aby byl příznak náležení změněn z původního na přiřazovaný.

Stejně tak v rámci indexů nového šuplíku je třeba provést zatřídění záznamu, které ovšem určuje operátor databáze a aplikace tedy jen zajistí správné zřetězení.

Jako v případě mazání může dojít k nutnosti přejmenovat celou sekci případně index. To nastává jen v případě, že je záznam vkládán jako první záznam indexu, nebo jako v případě mazání, opouští index, kde tvoří první a jediný záznam svého druhu.

Nový záznam

Vytvoření nového záznamu by naivním způsobem mohlo proběhnout stejně, jako se do kartotéky dostaly i všechny předchozí záznamy. To jest fyzickým napsáním záznamu, seskenováním a zařazením na příslušné místo.

I když by souběžně existovala ještě i výchozí lístková kartotéka a zadavatel by ji měl potřebu stále udržovat aktuální, je tento postup pořízení záznamu časově neefektivní a s výhledem do budoucna neefektivní i co do vztahu ke zpracovávaným datům. I přes pečlivé vedení nelze při duplicitní databázi zabránit chybám zaviněným lidským faktorem, proto tento systém bude používat patrně jen hodně málo zadavatelů.

Aplikace by měla umožnit ke každému záznamu, obsahujícímu jeden a více lístků, vést také strukturovaný databázový záznam. Taková existující struktura by umožnila časem záznamy přepsat a v ideálním případě plně digitalizovat a pomocí relací propojit.

I pokud by nedošlo k plné digitalizaci, výtěžnost databáze by se tím zvýšila.

V případě vytvoření nového záznamu by pak byl platný opačný postup, tj. v aplikaci by byly vyplněny příslušné strukturované atributy a na jejich základě by byl vytisknut záznam a zařazen do kartotéky.

Editace záznamu

Provádět změny v obrazové informaci je z pochopitelných důvodů obtížné.

Pro tento případ je nejvýhodnější postup Odstranění záznamu a následné vytvoření záznamu nového obsahujícího pozměněná data.

Vlastník databáze může trvat na zachování exitujícího obrazového záznamu.

Důvodem mohou být databází nepostihnutelná data, jako je například barva papíru nebo rukopis.

V takovém případě by aplikace měla umožnit vytvořit k záznamu záznam strukturovaný, existující obrazovou informaci doplnit o novou vygenerovanou obrazovou informaci, ale původní obrazová data zachovat, jako jeden z atributů záznamu.

Další funkce

Uživatelé lístkových kartoték se pravidelně potýkají problémem jak z určité skupiny záznamů vybrat skupinu a tu dále zpracovávat.

Jednou z doplňujících funkcí aplikace by měla být možnost s vybranými záznamy pracovat odděleně, dále je třídit, řadit, seskupovat atp.

3.7 Vyvíjení databázové aplikace

Pokud obrazová informace není pro uživatele nezbytná, je vyvinuta aplikace a navržena databáze na základě metod pro tvorbu databázových systémů. Tento text je zaměřen na digitalizovaných kartotékách postavených na skenovaných lístcích, proto se tématice databázových systémů nebudu na tomto místě více věnovat.

3.8 Plnění aplikace

Načtení automaticky předzpracovaných dat do aplikace je posledním bodem plánu. Uživateli i administrátorovi by měly být aplikací poskytnuty dostatečné nástroje k opravám a dalšímu rozvoji kartotéky.

Funkce v tomto bodě používané by měly být výsledkem analýzy stanovení cílů a měly by umožnit nejen spravovat data současná, ale i otvírat možnosti dalšího rozvoje.

Jsou to především aktivity směřující k plné strukturovanosti a propojenosti záznamů. Jako příklad uveďme možnost rozepsat záznam do atributů, možnost zavést číselníky a validovaná data, možnost tyto číselníky propojit se záznamy a propojit i záznamy mezi sebou.

Tento bod by neměl být nikdy fakticky uzavřen, pokud v kartotéce existuje potenciál ke zlepšení.

Kapitola 4

Systém RETROBI

4.1 Retrospektivní bibliografie v ÚČL AV ČR, v. v. i.

Retrospektivní bibliografie české literatury (RBČL), zpracovávaná více než 60 let v podobě lístkové kartotéky, tvoří jeden z nejcennějších sbírkových fondů Ústavu pro českou literaturu Akademie věd České republiky, v. v. i. Kartotéka je tvořena záznamy excerpcí článků z denního tisku a časopisů vycházejících v českém nebo německém jazyce na území českých zemí.

Jak excerpce probíhá? Excerptor vezme do ruky výtisk excerpovaného časopisu a pročítá jednotlivé články. Pokud se v článku nachází předmět zájmu, v případě RBČL například jméno literáta, literární věda, zmínka o literární činnosti nebo pokud byl článek přímo literátem sepsán, je o tomto článku vytvořen záznam. Záznam obsahuje údaje o autorovi článku, předmětu článku, údaje o zdroji, název novin nebo časopisu, stránku, případně krátkou citaci a některé další údaje.

Do excerpce bylo zahrnuto přibližně 530 titulů vycházejících v letech 1775-1945, sledovány byly pouze články dotýkající se literátů, literatury a literární vědy obecně.[1]

Potřeba řadit a vyhledávat podle více kritérií vedla autory k vytvoření několika katalogů, obsahujících prakticky shodná data, ovšem řazena jiným způsobem.

Informace o článku, který například napsal Arne Novák o Jaroslavu Vrchlickém byla zařazena jak do katalogu autorského k heslu „Novák, Arne“, tak

[1] Kartotéka tvoří uzavřený celek co do roků, ročníků, časopisů i témat. Proč tomu tak je a proč se dále nerozšiřuje?

Zpracování jednoho výtisku novin nebo časopisu popsaným způsobem je činnost náročná na pozornost, čas, odbornou způsobilost a ve výsledku tedy i na finanční prostředky. Získávání finančních prostředků pro vědecké účely je úzce navázáno na projekty s konkrétním výstupem. RBČL je možné označit za primární výzkum a jako takovému tyto výstupy chybí. Kartotéka je proto v této době považována za uzavřenou a doplňována pouze výjimečně.

do katalogu odkazovému k heslu „Vrchlický, Jaroslav“, i když fyzicky byl přes průklep záznam zapisován většinou jen jednou a po té zařazen.

Vznikla tak část autorská, řazena podle jmen autorů, návazná pomocná část dešifrační, obsahující excerpce se šifry autorů bez znalosti skutečného autora, části odkazová (předmětová-osobní), řazena podle předmětu excerpta a předmětová-věcná, řazená podle věcného obsahu s důrazem na předměty zájmu literární vědy. Následují části dešifrátová, řazena podle rozklíčovaných zkratek a k předešlým co do obsahu poměrně málo rozsáhlé části ikonické, snažící se postihnout grafický obsah jako jsou obrázky, fotografie, ilustrace atp.

Informace o jednom konkrétním článku se tak může nacházet hned v několika katalozích jedné kartotéky. Toto členění je samozřejmě velmi náročné na pečlivou redakční práci. Po každé excerpci bylo nutné pro jeden záznam vyhotovit hned několik kartotéčních lístků a ty správným způsobem zatřídit.

Několik desetiletí se, přes snahu zachovat původní koncept, podepsalo i na vývoji metodiky jak excerpta zpracovávat. Různá doba, různí excerptoři, různá představa o nepřesně definovaných oblastech zařazování (např. předmětová a ikonografická část). To je ovšem jen část dílčích problémů, se kterými se kartotéka potýká.

Kartotéka byla několikrát přetříděna, záznamy přehodnocovány a čištěny, metodiky zpřesňovány. I přes tuto péči a snahu zůstává stále více jak 13% autorských excerpt nedešifrováno a kartotéka stále podléhá procesu čištění, revizí a přetřiďování.

Výsledkem této práce je lístková kartotéka, čítající 1 700 000 excerpčních lístků. Z výše popsaných důvodů je počet skutečných jedinečných podchycených článků zhruba poloviční. Tento počet je však z hlediska literární vědy unikátní nejen v České republice, ale bez nadsázky lze napsat, že na celém světě.

4.2 Projekt RETROBI

S příchodem dostupné výpočetní techniky se logicky objevila myšlenka toto unikátní dílo převést do databázové podoby, a tak byl v devadesátých letech učiněn první pokus o strukturalizaci excerpt a jejich převod do počítače. Ačkoli se na projektu systematicky pracovalo více jak deset let, podařilo se do strukturované podoby převést pouze asi 90 000 záznamů, tedy méně jak 5% celkového objemu kartotéky.

Pomalý ruční přepis (30 let práce pro 1 osobu na plný úvazek[36]) a sílící tlak na zpřístupnění tohoto zdroje veřejnosti vedl k závěru, že zvolený systém práce není při daném způsobu finančního a personálního zabezpečení efektivní. Proto bylo rozhodnuto vypracovat projekt na digitalizaci kartotéky retrospektivní bibliografie skenováním.

V tříčlenném týmu jsme požádali o udělení grantu na tento projekt, který

byl posléze realizován v rámci programu INFOZ a za podpory MŠMT pod názvem „Digitalizace lístkového katalogu Retrospektivní bibliografie české literární vědy 1775-1945“ (VZ09004).

Realizaci celého projektu byla na straně ÚČL AV ČR v. v. i. zajistili tři lidé, kódování bylo svěřeno externí společnosti.

Mým úkolem byl dohled nad veškerými technickými aspekty projektu, tj. zpracování analýzy pro zadávací dokumentaci, dohled nad pracemi provedenými externím kodérem, zajištění potřebného hardwarového vybavení a následná správa systému. Projektu se dále účastnil odborník na RBČL, vedoucí Střediska literárněvědných informací (SLI) ve funkci hlavního koordinátora. Proces skenování zajistila zaměstnankyně oddělení SLI.

Výsledkem projektu byla nejen samotná digitalizace, ale zároveň software, který jsme nazvali RETROBI a který je použitelný i na obecnější typy kartoték.

Se záznamy nadále redakčně pracujeme a samotný systém RETROBI byl nasazen i na některé další kartotéky jiných institucí.

4.3 Příprava kartotéky pro digitalizaci

Než bylo skenované lístky možné přenést do databáze systému RETROBI, bylo třeba připravit je do jednotné podoby tak, aby samotný proces digitalizace běžel pokud možno co nejrychleji bez zbytečného zdržování.

4.3.1 Příprava a skenování šuplíků

Uživatelský interface systému RETROBI je optimalizován pro velikost kartotéčního lístku A6, přesto následující kroky zpracování jsou vhodné i pro jiné formáty.

4.3.1.1 Příprava vícestránkových lístků

Tvoří-li jeden záznam více lístků, jsou v RBČL spojeny v levém horním rohu sešívačkou. Tyto shluky by nebylo možné automaticky skenovat. Pro potřeby skenování proto byly rozešity a v šuplíku pootočeny o 90 °. V případě, že se takových záznamů nacházelo za sebou více, byla mezi ně vložena bílá papírová zarážka.

Takto oddělené lístky byly operátorem v průběhu skenování označeny jiným názvem souboru

4.3.1.2 Rozvržení indexů

Systém RETROBI již ve fázi skenování připravuje data pro dvouúrovňový vyhledavací strom sestávající v první úrovni z abecedy a v druhé úrovni z indexu podle primárního třídění kartotéky.

Rozvržení indexů vychází ze subjektivně „rozumně“ velké množiny záznamů, jejíž velikost byla po sérii pokusů stanovena na hodnotu mezi 200 - 300 lístky.

Tohoto členění je dosaženo umístěním 7-12 papírových zarážek na šuplík do míst kde vzniká abecední nebo logický předěl v množině záznamů.

Abecední předěl Na základní úrovni je to místo, kde končí písmeno A a začíná písmeno B, na nižších úrovních to může být místo, kde končí záznam „Drha“ a začíná záznam „Drtil“.

Logický předěl Toto dělení nemusí být, a v případě Retrospektivní bibliografie české literatury ani není, dostatečné. Určitá hesla kartotéky přesahují rozsahem svých záznamů i několik šuplíků a v takovém případě je třeba zvolit dodatečný klíč, který by umožnil vhodně nastavit indexy. V případě Retrospektivní bibliografie české literatury byl zvolen jako rozumný signifikátor předělu indexů rok publikování.

Index pak může například vypadat takto: „Vodák, Jindřich 1900“ a následující „Vodák, Jindřich 1910“.

Vzhledem k rozsahu a signifikantnosti roku publikování pro vědecký účel záznamu používá RBČL ještě čtvrtý pomocný identifikátor indexu, tím je pořadové číslo indexu s totožným rokem. Použijeme-li stejný příklad, to jest osobnost Jindřicha Vodáka, další skupina indexů vypadá takto: „Vodák, Jindřich 1926-1“ a následuje „Vodák, Jindřich 1926-2“.

Vocel	Vocílka	Vodák, Jindřich 1890
Vodák, Jindřich 1900	Vodák, Jindřich 1910	Vodák, Jindřich 1917
Vodák, Jindřich 1918	Vodák, Jindřich 1920	Vodák, Jindřich 1924
Vodák, Jindřich 1925	Vodák, Jindřich 1926-1	Vodák, Jindřich 1926-2
Vodák, Jindřich 1927-1	Vodák, Jindřich 1927-2	Vodák, Jindřich 1928-1
Vodák, Jindřich 1928-2	Vodák, Jindřich 1929-1	Vodák, Jindřich 1929-2
Vodák, Jindřich 1930-1	Vodák, Jindřich 1930-2	Vodák, Jindřich 1931-1
Vodák, Jindřich 1931-2	Vodák, Jindřich 1932-1	Vodák, Jindřich 1932-2
Vodák, Jindřich 1933-1	Vodák, Jindřich 1933-2	Vodák, Jindřich 1934-1
Vodák, Jindřich 1934-2	Vodák, Jindřich 1935	Vodehnal

[32]

Obrázek 4.1: Přehled indexů RBČL pro heslo „Vodák Jindřich“

Konkrétně tedy excerpta, jejichž autorem je v kartotéce RBČL Jindřich Vodák, čítají ohromujících 5531 záznamů, které jsou přehledně indexy rozděleny do 27 skupin. Každá ze skupin čítá v tomto případě například kolem 250 záznamů (poslední pochopitelně méně). A to i „plodná“ léta 1926-1934, která byla další úrovní dělení rozšířena o index roku. I několik tisíc záznamů je tímto jednoduchým způsobem zpřístupněno na několik málo pohybů po struktuře bez nutnosti složitě prohledávat kartotéku.

Předchozí kroky, i když je připraven podrobný návod jak je provádět, do jisté míry závisely na subjektivním posouzení obsahu kartotéky zkušeným vlastníkem. Všechny další kroky výsledek zobrazovaný systémem RETROBI mohou ovlivnit už jen minimálně.

4.3.2 Proces skenování

Lístky RBČL jsou psány nejen na standardním kancelářském papíru 80g/m^2, ale také na pauzovacím papíru, makulatuře, rozřezaných kartotéčních listech většího formátu, jednou větou na všem, co v průběhu šedesáti let bylo po ruce nebo k dispozici.

Výběru skeneru předcházel velmi podrobný průzkum trhu. Vypsal jsem nejprve čtyři poptávky, sledující dva zájmy rozdělené na dvě možnosti:

- Najít dodavatele skenovacího zařízení, které by bylo schopné oskenovat přibližně tři miliony stran.
- Najít dodavatele schopného oskenovat přibližně tři miliony stran za přijatelnou cenu.
- Najít dodavatele OCR technologie s maximální výtěžností.
- Najít dodavatele schopného provést OCR na třech milionech stran za přijatelnou cenu.

Výsledné nabídky bylo možné bohužel také rozdělit do dvou kategorií:

- Seriózní nabídky za neakceptovatelnou cenu.
- Nabídky, kde dodavatel nepochopil nebo neakceptoval zadání

Pro další hledání jsem se zaměřil odděleně pouze na hardware a jménem Akademie věd oslovil přímo větší společnosti, jmenovitě Xerox, Hawlett-Packard, Canon, Fujitsu-Siemens. Bohužel ani tato cesta nevedla k uspokojivému výsledku, pokud už firmy reagovaly, odkázaly mě na maloobchod.

Přesto se mi podařilo rozšířit osobní kontakty a po sérii dotazů na zařízení, jehož existenci a specifikaci jsem získával pouze z Internetu, jsem navštívil několik předváděcích center.

Hlavními kritérii pro výběr byly, v pořadí podle důležitosti, kvalita podavače (z důvodu široké škály používaného papíru), možnost oboustranného snímání, rychlost snímání, cena, počet doporučených kopií za týden a pochopitelně cena.

Pro zpracování RBČL byl vybrán skener Fujitsu-Siemens 6140, který při nastavení 600dpi a výkonu 20 snímků za minutu umožnil seskenovat oboustranně přibližně 3 000 000 stran jedním zaměstnancem během devíti měsíců.

Skener překročil limit snímků pro uznání záruky ještě před koncem pátého měsíce, nicméně je pravidelně čištěn a jsou mu měněna přítlačná kolečka podavače přibližně po 500 000 snímcích. Skener dodnes odvádí bezchybnou službu.

Soubory jsme skenovali z důvodu následného OCR v maximálním možném rozlišení, formátu BMP, 24 bitů.

4.3.3 OCR

Po zkušenostech s výběrem hardware jsem se u OCR software zaměřil na dva osvědčené a cenově dostupné produkty, Recognita a ABBY Fine Reader 9, u kterých jsem provedl sérii testů na významnějším vzorku dat.

Zvoleným programem se stal ABBY Fine Reader, verze 9.

Testy byly zaměřeny hlavně na získávání OCR záznamů ze souborů skenovaných na přístroji Fujitsu-Siemens 6140 do formátů JPG, PNG, TIFF a BMP, dále pak výsledky ze skenů obrázků v plném rozlišení formátu BMP, konvertovaných do JPG a PNG. Strojovým porovnáváním výsledných OCR jsem hledal rozdíly a v konkrétních lístcích následně srovnával se skutečným obsahem.

Tyto testy jsem opakoval v několika cyklech. Při použití menších rozlišení a nekompresních formátů vedlo k lepším výsledkům než větší rozlišení v kompresním formátu.

Pokusy ukázaly, že všechny varianty skenování mají vliv na pozdější OCR a i když procento rozdílů u skenů ve vysokém rozlišení bylo skutečně malé, po dohodě s vedoucím projektu jsme rozhodli skenovat ve formátu BMP, 24 bitů.

V tomto formátu bylo provedeno i OCR.

Následně jsme soubory zkonvertovali do PNG, protože tento formát dosáhl v OCR jen nepatrně horších výsledků, ale vedl k výrazné úspoře při komprimaci.

Proces OCR je časově velmi náročný, ale po nastavení bezobsluhového dávkového zpracování již nevyžaduje zásah operátora. Skenování a OCR proto probíhalo a probíhá dvoufázově - ve dne operátor skenuje, v noci nad daty spouští OCR.

4.3.4 Přejmenování, přiřazení OCR, první konverze

Jakmile je proces skenování a OCR dokončen, je použit program U1. Nástroj U1 zpracovává soubory na počítači operátora skenování ve třech krocích.

Nejprve jsou soubory přejmenovány do jednotného formátu. Poslední dvě číslice mají význam čísla lístku a strany lístku. Pokud tedy lístek končí např. xxxxx30.PNG znamená to, že je to třetí lístek záznamu a jeho první strana.

Jsou provedeny kontroly existence OCR souboru, pojmenování, logických návazností číslování a proběhne první konverze do souborů PNG v maximálním rozlišení. Tyto soubory jsou určeny k archivaci. Kontroly mají za cíl odstranit co nejvíce z chyb způsobených lidským faktorem.

Zároveň je provedena kontrola prázdných stran. Ta je založena na skutečnosti, že většina všech druhých stran lístků je prázdná. Bylo by tedy neefektivní se s nimi dále zabývat, nabízet je uživateli a zvětšovat jimi objem databáze. Program U1 spočítá index barevnosti stránky, najde její střední hodnotu a snaží se zjistit, zda počet výrazně tmavých bodů převyšuje zadané procento. Podle tohoto koeficientu vytipuje prázdné stránky a ty nabídne operátorovi skenování ke smazání. Je ovšem jen na něm, zda stránku skutečně smaže nebo v systému ponechá. Filosofie „jen člověk má právo v systému mazat“ se uplatňuje napříč celou aplikací.

4.3.5 Zmenšení, zařazení, upload

Pokud operátor skenování nenalezne žádnou závadu, je použit program U2. Tímto programem jsou soubory převzorkovány a zmenšeny. V databázi serverové části aplikace program zkontroluje existenci skupin podle názvu adresářů, případně je vytvoří a soubory do nich následně nahraje.

Pro potřeby zveřejnění na webu jsou původně skenované obrázky zbytečně velké, proto provádíme jejich konverzi. I tak jsou systémem RETROBI obrázky uvnitř aplikace ještě jednou zmenšovány a zdroj v nahrávaném rozlišení je uživateli poskytnut až v momentě, kdy mu zmenšeniny nestačí a požádá o maximální velikost.

Programy U1 a U2 jsou napsány v jazyku JAVA a jejich účelem je pouze připravit data a provést upload. Vzhledem k důležitosti jsou vybaveny mnoha ochrannými mechanismy proti opakovanému zapsání obrázků, proti přepsání dat a podobně. V dalším fungování digitální kartotéky již nemají žádný význam. Jednoúčelovost byla i argumentem externí firmy pro volbu technologie podle vlastního uvážení.

I když postoj firmy chápu, s tímto rozhodnutím nejsem spokojen a domnívám se, že použitá JAVA technologie je příliš závislá na verzi a v budoucnu může bránit nasazení produktu RETROBI.

4.4 Instalace aplikace RETROBI

Software pro digitalizaci kartotéky RBČL byl vyvíjen jako součást jednoho konkrétního projektu, pro tento konkrétní účel dolaďován, doprogramován a testován za provozu.

Konkrétní postup byl následující: Na hostitelském stroji jsem vytvořil virtuální server, na kterém koordinovaně a souběžně probíhalo několik akcí. Programátor externí společnosti vyvíjel a upravoval software podle našeho zadání a zároveň jsme do databáze za pomocí nástrojů U1 a U2 nahrávali oskenovananá data jejichž přesný formát jsme specifikovali již v době analyzování.

Formát uploadovaných dat se osvědčil. Během vývoje došlo pouze k několika situacím, kdy jsme byli nuceni celý postup nahrávání katalogu zopakovat od začátku z důvodů na straně software.

Domnívám se, že tento způsob práce výrazně urychlil realizaci celého projektu.

Nevýhoda tohoto postupu se projevila, když vznikla potřeba aplikaci použít pro jinou kartotéku a tedy znovu a jinde nainstalovat. Programátor externí společnosti po ukončení projektu předložil dokumentaci[35] s podrobným návodem jak instalaci v základních rysech zreplikovat, nicméně to zdaleka nestačí k bezproblémovému rozběhnutí aplikace.

I když jsem vytvořil několik pomocných instalačních skriptů, vzrůstající zájmem o projekt začal vyžadovat stále častější aktualizace. Po dohodě s vedením jsem do projektu přijal programátora, jehož úkolem bylo starat se o budoucí aplikace RETROBI a vytvořit automatický skript pro jejich instalaci. Skript bohužel ani po dvou měsících práce nebyl použitelný. Programátor byl nucen projekt opustit a skript byl z důvodu nefunkčnosti z projektu vyřazen. Nefunkční verze je k dispozici v adresáři `/src/retrobi.sh/fault` na přiloženém CD.

Na přiloženém CD je nyní k dispozici funkční instalační skript. Tento jsem vytvořil na základě původní nefunkční verze po prostudování programátorské dokumentace projektu, kódu projektu a dokumentace instalovaného software a aplikaci RETROBI je nyní možné bez nutnosti hlubšího studia celého systému nainstalovat a testovat.

4.4.1 Požadavky na hardware

Centrálním prvkem systému RETROBI je server, ke kterému se uživatelé připojují klienty HTTP. Serverová část byla odladěna pro operační systém Ubuntu Linux.

Doporučená minimální konfigurace je server se dvěma procesorovými jádry, 2GiB RAM, diskové pole 2TiB (pro přibližně 1,5 milionu lístků).

Výkon celého systému závisí především na přístupech k pevnému disku, proto je pro vyšší spolehlivost a odolnost doporučena technologie RAID.

4.4.2 Instalace systému

Software RETROBI je nástroj velmi citlivý na konfiguraci operačního systému. K bezchybné činnosti potřebuje několik desítek balíčků a knihoven. Instalujte jen na stroj dedikovaný pouze pro potřeby RETROBI. Úspěšná instalace je očekávána na server s holou instalací operačního systému.

Instalační skript jsem odladil pro holou instalaci operačního systému Linux Ubuntu Server 14.04.1 LTS[14].

Pro automatickou instalaci postupujte podle následujícího schématu:

1. Z přiloženého CD nakopírujte adresář `src` do home uživatele.
2. Na server se přihlaste jako uživatel. Automatická instalace probíhá pod standardním uživatelským účtem.

V žádném případě nespouštějte instalaci pod uživatelem **`root`**, následně by docházelo k chybě způsobené uživatelskými právy. Ostatně automatický instalační skript spuštění pod uživatelem **`root`** ani nepovolí.

3. Spusťte instalační skript **`install.sh`** v adresáři **`~/src/retrobi.sh`**.

4. Po skončení skriptu zadejte do internetového prohlížeče adresu `http://doménové_jméno_serveru/test`

Na začátku jsou provedeny standardní kontroly, zda již instalace není spuštěna a je proveden i test, pod jakým uživatelem byla spuštěna.

Vřele doporučuji instalaci provádět na holou čerstvou instalaci, nicméně je možné, že bude třeba instalaci zopakovat a v tom případě není bezpodmínečně nutné instalovat operační systém serveru znovu.

Instalační skript v následujících krocích provede pokus o odstranění důsledků předešlé instalace. Pokusí se zastavit databázi CouchDB[3], servery Tomcat[5] a Apache[4], dále se pokusí odinstalovat balíčky, kdyby již byly přítomny.

V následujících krocích proběhne samotná instalace nezbytných balíčků. Kromě jiných je to

- Apache HTTP Server[4],
- Apache Tomcat[5],
- Apache CouchDB[3],

a řada dalších.

Instalační skript načte proměnné ze souboru **`settings.sh`** a v dočasné složce připraví nastavení konfiguračních souborů, která následně překopíruje na jejich originální lokaci.

Po zdařilé instalaci všech částí RETROBI jsou spuštěni potřební démoni.

Dále jsou nahrána ukázková data a spouštěny databáze. I když ukázková data nejsou příliš rozsáhlá, proces načítání a spouštění databáze je časově náročný a je třeba vyčkat dostatečně dlouhou dobu.

I po skončení skriptu je třeba vyčkat než je do webového prohlížeče zadána adresa. Není-li databáze ve chvíli volání připravena, zobrazí se hláška „Probíhá údržba databáze“, v opačném případě se objeví úvodní stránka nainstalovaného produktu s ukázkovými daty[2].

Defaultní přihlášení do aplikace je kombinací jména a hesla **`admin:admin`**.

Aplikace je nyní připravena na nahrávání dat přes nástroje U1 a U2.

[2]Testovací verze obsahuje 24 záznamů rozdělených pod písmena A, B a C. Data jsou převzata z lístkového katalogu „Staročeská sbírka“[7] s laskavým svolením Mgr. Vojtěcha Malínka, Ph.D.

Přestože skript projde důležité kroky, které by jinak bylo třeba provést manuálně, automatická kompilace ve verzi JDK 7 neprobíhá spolehlivě a v souladu s dokumentací.

Pro instalační skript jsem proto použil některé předkompilované části, které je možné v případě zájmu nahradit vlastní, ručně zkompilovanou, verzí.

Kompilace se spustí příkazem `mvn` v příslušném adresáři. V něm vznikne adresář `target` a objeví se nový soubor s příponou `.war`. Jeho překopírováním do složky `webapps` serveru Tomcat je kompilace dokončena.

Ve složce `/opt/RETROBI/retrobisrc` jsou nahrány zdrojové kódy projektu RETROBI.

4.4.3 Použité technologie

Projekt RETROBI je naprogramován v objektově orientovaném jazyce Java (verze 1.6), některé části pak ve skriptovacím jazyce JavaScript.

Pro uložení dat je použita dokumentově orientovaná NoSQL databáze CouchDB[3].

Webová aplikace je postavena na frameworku Apache Wicket[6], produkční verze běží v kontejneru Tomcat[5].

Pro fulltextové vyhledávání je použit servlet CouchDB Lucene.

4.5 Zálohování dat

4.5.1 Zálohování zdrojů

Před zahájením skenování jsem provedl sérii pokusů na datech. Oskenováním jednoho šuplíku jsem získal přibližně tisíc záznamů obsahujících průřez různorodých lístků, co do barevnosti a hustoty písma. Následně jsem provedl sérii pokusů zaměřených na OCR a i na kompresní technologie, především zip a rar.

Z výsledků vyplynulo, že největší úspory místa je dosaženo, budeme-li uchovávat skenované obrazy ve formátu PNG, v původní velikosti, komprimované programem 7-Zip.

Pro potřeby další manipulace jsme zvolili kompresi po jednotlivých indexech s přibalenou informací původního názvu souboru. Data jsou následně nahrána na obyčejný pevný disk připojený k počítači HDD dokovací stanicí. Druhá záloha vzniká okopírováním disku, obě zálohy jsou uloženy na geograficky vzdálených místech.

Tímto řešením jsme dosáhli nejlepšího poměru výkon/cena při uchování jedné zálohy na externím úložišti.

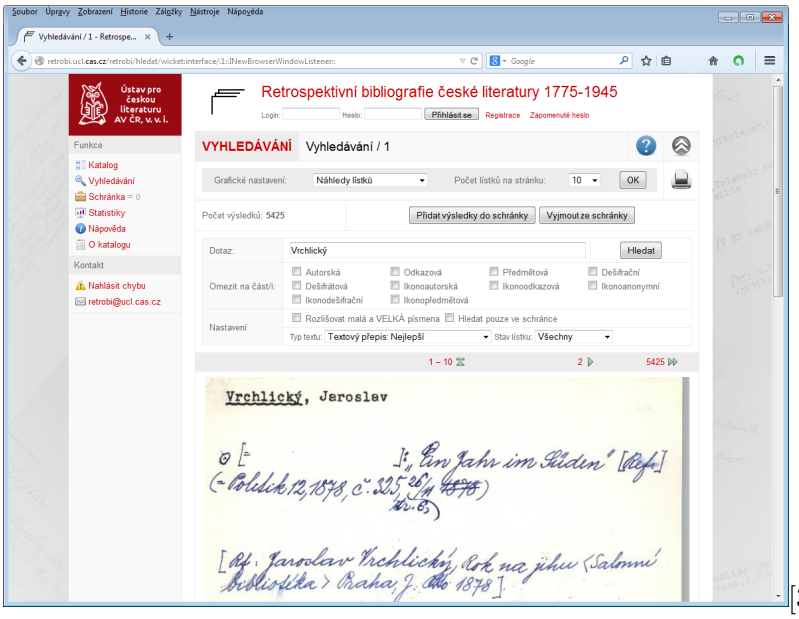
Uchování dat v této podobě je pro chod samotné kartotéky bezvýznamné. RETROBI obsahuje redukované obrazy a je zálohováno jiným způsobem. Ovšem skeny v plném rozlišení mohou mít velký datový potenciál do budoucnosti.

Vzhledem k tomu, že nové oskenování celé kartotéky v plném rozsahu není pravděpodobné, jedná se o zdroj, který může při významnějším posunu v oblasti OCR poskytnout další cenná data.

4.5.2 Zálohování databáze a aplikace

Počáteční konfigurace systému RETROBI není zcela triviální činností a vzhledem k velikosti databáze RBČL je velikost operačního systému zanedbatelná. ÚČL AV ČR, v. v. i. většinu svých strojů virtualizuje a proto zde zálohování probíhá formou zálohy celého virtuální stroje.

4.6 RETROBI z pohledu obecné analýzy



[32]

Obrázek 4.2: Systém RETROBI, ÚČL AV ČR v. v. i.

Digitalizaci kartotéky RBČL předcházela několikaměsíční analýza. V první fázi bylo nutné, abych pochopil, jakým způsobem fungují kartotéky obecně a jaká jsou specifika RBČL. Zároveň bylo nezbytné, abych naopak zbylým členům týmu vysvětlil jak, alespoň v principu, fungují databáze, jaké jsou jejich možnosti, omezení, co je v našich silách naplnit.

Tuto vizi bylo třeba vměstnat do časově i finančně omezeného rámce projektu. Vznikla velmi podrobná zadávací dokumentace, na jejímž základě bylo vypsáno výběrové řízení na externího dodavatele, který provede kódování celého projektu. Ačkoli bylo přizváno několik desítek firem, seriózní plán předložily jen dvě.

Nabízená řešení byla shodou okolností značně protichůdná. První společnost nabídla řešení na základě klasického SQL modelu, druhá novou technologii NoSQL.

Na základě výběrového řízení, ve kterém hlasovala komise složená z vedení ÚČL AV ČR, v. v. i., zvítězila společnost nabízející NoSQL technologii dokumentově orientované databáze CouchDB.

Řešení NoSQL jistě přineslo řadu výhod při kódování, jako je například možnost soustředit se na aplikaci a neřešit různorodost atributů. I když tento aspekt přispěl k urychlení celého vývoje, z mého pohledu toto řešení není šťastné. Kartotéka RBČL obsahuje množství relací, které jsou vysledovatelné, ale díky technologii NoSQL nepodchytitelné. Na tento rozpor jsem upozorňoval již při samotném výběrovém řízení a logicky na něj došlo i v průběhu kódování celého projektu, kdy vznikl spor o naplnění jednoho z bodů analýzy obsažené v zadávací dokumentaci. Je samozřejmě otázkou diskuse, do jaké míry může být relačnost tak rozsáhlé kartotéky budována. To byl jeden z hlavních argumentů vítězné firmy pro obhajobu postoje relace nezavádět. Ovšem budeme-li nahlížet na RETROBI ne jako na jednoúčelový projekt, ale jako na řešení aspirující na označení „univerzální“, je to důležitý rys, který chybí.

4.7 RETROBI z pohledu uživatele

Srovnáme-li řešení RETROBI s jinými českými existujícími online kartotékami, najdeme mnoho společných prvků a principy, které jsou obsaženy i v metodice RETROKON[24].

Funkčností a ergonomií pohybu po záznamech ale výrazně převyšuje všechny aplikace uvedené v kapitole 2. Na rozdíl od technologie RETROKON řeší RETROBI i další vytěžování a uchovávání dat v rámci jedné aplikace.

Naplňuje primární cíl, to jest zpřístupnění katalogu prostřednictvím Internetu, ale zároveň také umožňuje rozpracování převodu kartotéky do několika etap takovým způsobem, aby kartotéka i v době strukturování nebo přepisu mohla plnohodnotně fungovat.

Řešení RETROBI poskytuje možnost postupovat v digitalizaci od obrazů až po plnou databázovou podobu (s výjimkou výše zmíněné nerelačnosti). V několika stupních je možné po obrazu využít OCR pro hypertextové prohledávání, tento text použít jako základ pro čistopis, dále pak text segmentovat na základní bloky a v nejvyšším stupni data rozčlenit na jednotlivé atributy.

Opačný postup je víceméně shodný, ale strojově zpracovatelný. Po zadání atributů se vytvoří segmentovaný záznam, přepis celého lístku a vygeneruje se i zpětně obrazová podoba.

Databáze tak umožní vyhledávat na všech stupních, ať už holým procházením obrázků nebo s přídavkem textového vyhledávání na různých úrovních.

Uživatelský komfort je dále doplněn řadou rešeršních funkcí jakými je uklá-

dání lístku do schránky, tvoření vlastních seznamů a vlastního řazení, vlastních poznámek k lístku a mnoho dalších užitečných funkcí.

Popis všech funkcionalit uživatelského rozhraní by zabral mnoho stran a lze se o nich dočíst v nápovědě RETROBI[37]. Z mnoha funkcí administračního rozhraní stojí za zmínku rozhraní pro nakládání s uživatelskými opravami. Uživatelé mohou nejen hlásit zjištěné chyby, ale také se mohou aktivně zapojit do přepisu záznamů na všech úrovních - opravou OCR, segmentací i rozepsáním atributů. Výsledek je okamžitě k dispozici všem uživatelům, ale do databáze je zařazen až po schválení administrátorem. Aby se zabránilo vtipálkům a škůdcům narušovat touto cestou integritu databáze, je funkce povolena pouze registrovaným uživatelům.

V této chvíli je systém RETROBI nasazen v Ústavu pro českou literaturu AV ČR, v. v. i., pro projekty „Retrospektivní bibliografie české literatury 1775-1945“ a „Staročeská sbírka“, dále pak v Institutu umění Divadelního ústavu pro projekt „Retrospektivní bibliografie českého divadla“ a připravují se další nasazení.

Systém je postavený na licenčních ujednáních AGPL, otevřený každému, kdo by ho chtěl doplnit nebo přizpůsobit svému záměru.

Kapitola 5

Bezpečnost aplikace RETROBI

Na otázku zabezpečení aplikace RETROBI můžeme nahlížet z mnoha pohledů. Jedním z nich je nebezpečí neoprávněného vytěžení a zneužití dat v databázi obsažených, další je cílení na stabilitu a dostupnost databáze, šíření malware, spamu a další hrozby.

Lístkové kartotéky jsou nabízeny uživatelům na Internetu zdarma. Riziko, že jejich obsah bude použit pro neoprávněnou manipulaci je z tohoto pohledu minimální. Budu tedy nadále předpokládat, že aplikace RETROBI může být ohrožena především útokem na aplikační část.

Zajistit obecnou bezpečnost aplikace je velmi náročný úkol.

Potenciální útočník může využít velkou řadu cest jak na cíl zaútočit a proto i starost o zabezpečení je velmi komplexní, stále živá a stále se vyvíjející činnost.

Útočný vektor může směřovat na různé úrovně systému.

Může být veden jak na fyzický, případně virtuální, server, na kterém je aplikace spuštěna, ale rovněž může cílit na samotný webový server nebo kontejner poskytující data a v neposlední řadě též na samotnou aplikaci.

Ochrana před každou z těchto možností útoku musí být umístěna vždy na jiném místě topologie.

Úroveň síťového zabezpečení je umístěna na IP firewallu, který je většinou řešen jako samostatný fyzický stroj popřípadě jako softwarové řešení na serveru samém.

Zabezpečení software webového serveru se řeší spuštěním pouze s nezbytnými právy, často pomocí sandboxingu. Příkladem tohoto způsobu je například SELinux[18] či AppArmor[17] v OS Linux.

Na tomto místě nás bude zajímat pouze případ zabezpečení aplikace jako takové.

To se děje navíc také na nejvyšší sedmé aplikační vrstvě OSI modelu, opět buď na nějakém jiném virtuálním či fyzickém stroji, fungujícím zároveň jako reverzní proxy server, v modulu software konkrétního webového serveru, nebo rovněž formou cloudové služby. Aplikační firewall má, na rozdíl od IP

firewallu, komplexnější přehled o protékajících datech. Například kromě cílové a zdrojové adresy může rovněž zkoumat i adresu požadované webové stránky.

Aplikační firewally obecně kontrolují nejen vstup a přístup ale rovněž výstup. Pomocí kontroly výstupu se dají odhalit jinak neodhalené útoky, či zabránit únikům citlivých dat. Například je samozřejmě nežádoucí, aby vnitřní síť či server opustila databáze se jmény uživatelů nebo zákazníků, zdrojové kódy aplikace či nějaký jiný důvěrný dokument.

Aplikační firewall je řízen pravidly. Tato pravidla nám definují obecné vzorce nežádoucího provozu. V tomto případě je cílem analyzovat HTTP či HTTPs provoz jakožto typických webových protokolů. Tyto protokoly mají definovaný obsah paketů a jednotlivé příznaky spojení jako je např. SYN, FIN a další pro jednotlivé fáze komunikace. Jedním ze základních pravidel může být například ochrana před SYN-flood DoS (Denial Of Service) útokem. Při běžné komunikaci zašle klient serveru požadavek SYN a ten mu odpoví. Pokud těchto požadavků přichází velké množství, tak se pravděpodobně jedná o DoS útok. Těchto požadavků může chodit velké množství z velkého množství adres typicky z nějakého botnetu použitého k distibuovanému útoku neboli DDoS (Distributed Denial Of Service). Tento typ útoku by ještě mohl být zachycen stavovým firewallem, který zkoumá i vyšší než třetí vrstvu OSI modelu, ale část produkčních IP firewallů je stále ještě bezstavová a tedy proti těmto útokům bezbranná, a proto je vhodné je mít ošetřeny i v aplikačním firewallu.

Nejčastější webová bezpečnostní rizika jsou popsána na webu projektu OWASP (Open Web Application Security Project) `https://www.owasp.org`, který sdružuje zájemce a odborníky o webovou aplikační bezpečnost. Mezi tato rizika se dlouhodobě zařazují například SQL-injection, XSS (Cross-Site Scripting), CSRF (Cross-Site Request Forgery) a mnoho dalších.

Jako velmi jednoduchý ukázkový příklad se nejčastěji uvádí útok na nezabezpečený vstup od uživatele. Do aplikace je možné se přihlásit uživatelským jménem a heslem. Ale místo jména se tam vyplní pouze ' or '1'='1. Tento příkaz je následně vyhodnocen SQL serverem jako správný, protože v logické podmínce nebo je druhá možnost pravdivá tedy jedna se opravdu rovná jedna. V této chvíli je útočník přihlášen do systému bez znalosti jména či hesla. Jakkoli tento příklad působí velmi úsměvně, tak je to největší riziko mnoha aplikací.

Aplikace bývají v dnešní době čím dál složitější a proto se používá více a více externích knihoven či frameworků. Jako příklad bych uvedl bohužel velmi časté chyby v různých CMS typu WordPress, Drupal či Joomla, které jsou často doplněné různými moduly. V takto komplikovaných strukturách se poté těžko hledají možná rizika a proto je výhodné nasadit webový aplikační firewall.

S rostoucím použitím různých webových služeb roste množství potenciálních cílů respektive uživatelů. Na chyby v nejrůznějších místech aplikací cílí množství různého malware a virů. Tempo jejich obměn je obrovské a dopředu zabezpečit aplikaci je nemožné.

Velké množství těchto rizik má podobný scénář a tedy se dají najít určité vzory a pomocí nich zachytit i nové či netypické útoky.

5.1 Funkce mod_security

Nyní se zaměřím pouze na metodu ochrany webového serveru s aplikací RETROBI pomocí vloženého bezpečnostního modulu, v našem případě se jedná o webserver Apache a modul mod_security[19].

Samotný bezpečnostní modul se dá přiřadit i přímo do aplikace RETROBI či do aplikačního kontejneru Tomcat spuštěného v Apache Tomcat, je to vlastně další aplikace do Tomcatu.

Příchozí požadavek na server je postupně zpracováván různými procesy a moduly, a po té předán na vyřízení podle typu požadavku buď nějaké další službě např. Tomcat či zpracován již tímto softwarem. Do tohoto procesu po přijetí vstupuje mod_security a podle definovaných pravidel buď požadavek propustí, zaznamená, zamítne nebo provede kombinaci těchto možností.

Ve výchozím nastavení modul pouze zaznamenává podezřelý provoz a vše je dohledatelné v logách a dle nich je pak možné upravit jednotlivá pravidla podle konkrétních výsledků.

5.2 Instalace a aktivace mod_security

Všechny zde uváděné příklady jsem testoval na serveru retrodipl.ucl.cas.cz s IP adresou 147.231.80.132, sken byl proveden ze serveru s IP 147.231.80.88.

Referenční instalaci jsem provedl ve formě zásuvného modulu do serveru Apache2, který v mé instalaci funguje jako proxy pro kontejner RETROBI běžící v Tomcatu. Použiji nejobvyklejší metodu instalace z předpřipravených balíčků, aby při případném update nebylo potřeba provádět rekompilaci.

Celý modul je možno také zkompilovat přímo ze zdrojových kódů, ale z hlediska dlouhodobé údržby, a tedy i bezpečnosti, se jeví lépe použití balíčku.

Samotná instalace se provede příkazem:

```
# apt-get install libapache2-modsecurity
```

Nyní mám nainstalovaný modul včetně ukázkových pravidel.

Samotná pravidla je možno stáhnout i přímo z webového odkazu `https://github.com/SpiderLabs/owasp-modsecurity-crs/archive/master.zip` či pomocí systému správy „git“, ale verze pravidel v balíčku by měla být pro prevenci většiny útoků dostačující.

Modul je sice nainstalovaný, ale je nutné aktivovat konfiguraci následujícími příkazy

```
# mv /etc/modsecurity/modsecurity.conf{-recommended,}
```

```
# service apache2 reload
```

Nyní již můžeme přistoupit k samotné konfiguraci.

Nejdůležitější volba se nazývá „SecRuleEngine“, ta nám udává vlastní chování modulu.

Ve výchozím stavu je nastavena na hodnotu „DetectionOnly“.

V tomto módu pouze loguje bezpečnostní incidenty a funguje tedy jako IPS, který jen upozorní na možné hrozby. V tomto nastavení by se měly sledovat výstupy nacházející se v souboru

```
# /var/log/apache2/modsec_audit.log
```

a dle toho se rozhodnout, jaká pravidla následně aktivovat.

Až po té, co bude vše ozkoušené, je možné změnit volbu „SecRuleEngine“ na hodnotu „On“ a pečlivě dále sledovat provoz, aby se omylem web neznepřístupnil.

Volbu jednotlivých pravidel nastavím editací souboru

```
/etc/apache2/mods-available/security2.conf
```

kam umístím directivu

```
include "/usr/share/modsecurity-crs/activated_rules/*.conf"
```

mezi počáteční a koncový tag.

V adresáři **activated_rules** mám nyní nastavenou aktivaci všech pravidel.

Jednotlivá pravidla aktivuji podobně jako stránky či moduly v serveru Apache vytvořením symbolického odkazu do zvoleného adresáře. Je možné například aktivovat ochranu proti obecným útokům typu spouštění systémových procesů, chybnému rozdělování http požadavků a dalších tímto příkazem

```
# ln -s /usr/share/modsecurity-crs/base_rules \
/modsecurity_crs_40_generic_attacks.conf \
/usr/share/modsecurity-crs/activated_rules
```

Pravidla jsou rozdělena do několika kategorií.

Nejčastěji jsem používal pravidla v adresáři **base_rules**, kde se nachází vzory pro rozpoznání XSS, neboli cross-site scripting útoků, vkládání kódu pomocí SQL-injection, skenování či chybných požadavků podle http standardů. Další kategorií jsou experimental rules, kde se nachází typické vzorce útočných vektorů útoků DoS, lámání hesel metodou brute-force a dalších. V adresáři **optional_rules** se pak nachází pravidla například proti vkládání spamových příspěvků do diskuzí a mnoho dalších.

Toto jsou samozřejmě pouze ukázková pravidla od organizace OWASP. Na internetu můžeme nalézt další kolekce pravidel od jiných společností. Další možností je vytvoření vlastních pravidel pro potřeby konkrétní implementace.

5.3 Testování a výstup logu

Nyní mám nakonfigurovaný modul mod_security a aktivovaná základní pravidla prevence útoků. Všechna zaznamenaná porušení pravidel nalezneme v logu v souboru

```
# /var/log/apache2/modsec_audit.log
```

Pokud máme server přístupný z Internetu, tak se nám logy poměrně rychle začnou plnit záznamy o skenování a dalších automatických útocích na různý software či aplikace. Abych si to mohl sám ověřit, tak využiji některý z mnoha dostupných skenerů zranitelností v tomto případě „nikto“[20]. Na jiném serveru si spustím obecný sken základních zranitelností

```
# nikto -host retrodipl.ucl.cas.cz
```

Na obrázku 5.1 můžeme sledovat průběh výsledků testování. V logu jsou zachyceny adresy útočícího stroje s IP 147.231.80.88 a testovaného stroje s instalací aplikace RETROBI s IP 147.231.80.132. Log je zaplněn varovnými hláškami s odkazy na soubory pravidel, která byla porušena.

V dalším testování přepnu modul mod_security do aktivního módu, kdy už působí podobně jako IPS a znovu spustím testování.

Z výpisu logu na obrázku 5.2 je patrné, že požadavek byl zamítnut po vyhodnocení první žádosti v rozporu s pravidly. Útočník tak byl v tomto případě odražen, testováním dalších pravidel už v takovém případě nemá smysl.

Podrobným zkoumáním dalších výstupů skeneru „nikto“ se dá vysledovat 34 potencionálních nalezených zranitelností při zapnutém pasivním logování a po zapnutí již skener nenašel zranitelnosti žádné. Tento výstup je uložen do souboru **`modsec_audit.log`** na přiloženém CD.

Zajištění komplexní bezpečnosti je velmi složitá záležitost a zde uvedený postup je pouze triviální úvod do problematiky. I tak je však vidět, že tento postup dokáže odrazit mnoho automatizovaných útoků a také zkomplikovat život případnému útočníkovi. Po zapnutí bezpečnostního modulu do aktivního módu je třeba pečlivě otestovat, zda nedochází k nežádoucímu blokování legitimních požadavků.

```
--1d1b9e7b-A--
[03/Dec/2014:16:24:29 +0100] VH8rLX8AAQEAAExs9YYAAAAC 147.231.80.88 37045 147.231.80.132 80
--1d1b9e7b-B--
HEAD / HTTP/1.1
Connection: Keep-Alive
User-Agent: Mozilla/4.75 (Nikto/2.1.4) (Evasions:None) (Test:Port Check)
Host: retrodipl.ucl.cas.cz

--1d1b9e7b-F--
HTTP/1.1 200 OK
Keep-Alive: timeout=5, max=100
Connection: Keep-Alive
Content-Type: text/html;charset=UTF-8

--1d1b9e7b-E--

--1d1b9e7b-H--
Message: Warning. Operator EQ matched 0 at REQUEST_HEADERS. [file "/usr/share/modsecurity-crs/activated_rules/modsecurity_crs_21_protocol_anomalies.conf"]
 [line "47"] [id "960015"] [rev "1"] [msg "Request Missing an Accept Header"] [severity "NOTICE"] [ver "OWASP_CRS/2.2.8"] [maturity "9"] [accuracy "9"] [t
ag "OWASP_CRS/PROTOCOL_VIOLATION/MISSING_HEADER_ACCEPT"] [tag "WASCTC/WASC-21"] [tag "OWASP_TOP_10/A7"] [tag "PCI/6.5.10"]
Message: Warning. Matched phrase "nikto" at REQUEST_HEADERS:User-Agent. [file "/usr/share/modsecurity-crs/activated_rules/modsecurity_crs_35_bad_robots.co
nf"] [line "20"] [id "990002"] [rev "2"] [msg "Request Indicates a Security Scanner Scanned the Site"] [data "mozilla/4.75 (nikto/2.1.4) (evasions:none) (
test:port check)"] [severity "CRITICAL"] [ver "OWASP_CRS/2.2.8"] [maturity "9"] [accuracy "9"] [tag "OWASP_CRS/AUTOMATION/SECURITY_SCANNER"] [tag "WASCTC/
WASC-21"] [tag "OWASP_TOP_10/A7"] [tag "PCI/6.5.10"]
Message: Warning. Pattern match "(?i:(?:c(?:o(?:n(?:t(?:entsmartz|actbot/)|cealed defense|veracrawler)|mpatible(?: ;(?: msie|\\.)|-)|py(?:rightcheck|guard
)|re-project/1.0)|h(?:ina(?: local browse 2\\.|claw)|e(?:rrypicker|esebot))|rescent internet toolpak)|w(?:e(?:b(?: (?:downloader|by ..." at REQUEST_HEADER
S:User-Agent. [file "/usr/share/modsecurity-crs/activated_rules/modsecurity_crs_35_bad_robots.conf"] [line "27"] [id "990012"] [rev "2"] [msg "Rogue web s
ite crawler"] [data "Nikto"] [severity "WARNING"] [ver "OWASP_CRS/2.2.8"] [maturity "9"] [accuracy "9"] [tag "OWASP_CRS/AUTOMATION/MALICIOUS"] [tag "WASCT
C/WASC-21"] [tag "OWASP_TOP_10/A7"] [tag "PCI/6.5.10"]
Message: Warning. Operator GE matched 5 at TX:inbound_anomaly_score. [file "/usr/share/modsecurity-crs/activated_rules/modsecurity_crs_60_correlation.conf
"] [line "37"] [id "981204"] [msg "Inbound Anomaly Score Exceeded (Total Inbound Score: 10, SQLi=0, XSS=0): Rogue web site crawler"]
Apache-Handler: httpd/unix-directory
Stopwatch: 1417620269145714 89033 (- - -)
Stopwatch2: 1417620269145714 89033; combined=80642, p1=77423, p2=1927, p3=53, p4=183, p5=597, sr=90, sw=459, l=0, gc=0
Response-Body-Transformed: Dechunked
Producer: ModSecurity for Apache/2.7.7 (http://www.modsecurity.org/); OWASP_CRS/2.2.8.
Server: Apache/2.4.7 (Ubuntu)
Engine-Mode: "DETECTION_ONLY"

--1d1b9e7b-Z--
```

Obrázek 5.1: Ukázka probíhajícího útoku, mod_security v módu pasivního logování „DETECTION ONLY“

```
--1fd1197e-A--
[03/Dec/2014:16:26:41 +0100] VH8rsX8AAQEAAEzcRakAAAABB 147.231.80.88 37141 147.231.80.132 80
--1fd1197e-B--
HEAD / HTTP/1.1
Connection: Keep-Alive
User-Agent: Mozilla/4.75 (Nikto/2.1.4) (Evasions:None) (Test:Port Check)
Host: retrodipl.ucl.cas.cz

--1fd1197e-F--
HTTP/1.1 403 Forbidden
Keep-Alive: timeout=5, max=100
Connection: Keep-Alive
Content-Type: text/html; charset=iso-8859-1

--1fd1197e-E--

--1fd1197e-H--
Message: Access denied with code 403 (phase 2). Operator EQ matched 0 at REQUEST_HEADERS. [file "/usr/share/modsecurity-crs/activated_rules/modsecurity_cr
s_21_protocol_anomalies.conf"] [line "47"] [id "960015"] [rev "1"] [msg "Request Missing an Accept Header"] [severity "NOTICE"] [ver "OWASP_CRS/2.2.8"] [m
aturity "9"] [accuracy "9"] [tag "OWASP_CRS/PROTOCOL_VIOLATION/MISSING_HEADER_ACCEPT"] [tag "WASCTC/WASC-21"] [tag "OWASP_TOP_10/A7"] [tag "PCI/6.5.10"]
Action: Intercepted (phase 2)
Stopwatch: 1417620401886510 8133 (- - -)
Stopwatch2: 1417620401886510 8133; combined=2288, p1=1840, p2=228, p3=0, p4=0, p5=219, sr=236, sw=1, l=0, gc=0
Response-Body-Transformed: Dechunked
Producer: ModSecurity for Apache/2.7.7 (http://www.modsecurity.org/); OWASP_CRS/2.2.8.
Server: Apache/2.4.7 (Ubuntu)
Engine-Mode: "ENABLED"

--1fd1197e-Z--
```

Obrázek 5.2: Ukázka zamítnutí požadavku útočníka, mod_security v aktivním módu „ENABLED“

Závěr

Stanovené cíle

„Seznamte se s problematikou převodu lístkových kartoték do digitální - databázové podoby“

S lístkovými kartotékami se setkávám již tři desetiletí, nejprve v pozici uživatele a v posledních patnácti letech i z pozice správce a autora při své práci v ÚČL AV ČR, v. v. i. Při studiu materiálů pro tuto práci jsem získal větší přehled o situaci na poli této problematiky v České republice i ve světě.

„Srovnejte existující řešení digitalizace lístkových kartoték“

Srovnání je provedeno porovnáním filosofie a funkčnosti několika konkrétních instalací v kapitole 2 od strany 5.

„Vytvořte základní analýzy pro převod lístkových kartoték do digitální podoby“

Této problematice se věnuje kapitola 3 od strany 17.

„Porovnejte svůj návrh s existujícími řešeními a systémem RETROBI“

Postupy digitalizace a nastavení je věnována kapitola 4 od strany 29, rozdílům ve funkčnosti je pak samostatně věnován oddíl 4.6, strana 39.

„Navrhněte sadu nástrojů k dosažení automatizace instalace a konfigurace systému RETROBI“

Celý proces instalace a konfigurace vyžaduje nejen pokročilejší znalosti administrace systému Linux, ale také pochopení principů, na kterých je aplikace RETROBI postavena.

Integrováním všech důležitých částí instalace do jednoho skriptu se snažím komplikované fáze tohoto procesu překlenout. Zájemce získává nástroj, který

mu umožní zkrátit studium manuálů na minimum a začít se hned prakticky seznamovat s produktem.

Proces zjednodušené instalace je popsán v oddílu 4.4, strana 35

„Prozkoumejte možnosti využití web application firewallů a nakonfigurujte ukázkové řešení, ověřte pro systém RETROBI“

Nasazení aplikace RETROBI do prostředí s veřejným přístupem je pro využitelnost kartotéky uživateli nezbytné. Rizika s tím spojená jsou rozebírána v kapitole 5 od strany 43, nasazení aplikačních firewallů se věnuje pododdíl 5.1 od strany 45.

Shrnutí

Problematika převodu lístkových kartoték je velmi obsáhlá a pro různé kartotéky specifická. Ostrá hranice přechodu mezi záznamem v reálném světě a jeho obrazem ve světě digitálním znamená velký zlom i ve způsobu náhledu na záznam a na jeho relace k ostatním záznamům. Pohled knihovníka na kartotéku se zásadně liší od pohledu programátora a i když je tento pohled do určité míry dán generačními rozdíly, vzdálenost oborů informatiky a knihovnictví je i tak značná.

V posledních dvaceti letech bylo v České republice zpracováno několik projektů, metodik a vytvořen různý software pro digitalizaci a prezentaci kartoték. I když se v závislosti na míře abstrakce projekty od sebe liší, základ vychází u všech ze stejného principu a výsledek se dá rozlišit hlavně mírou uživatelského komfortu a, co je z mého pohledu nejdůležitější, možností dalšího rozvoje.

Nástroj RETROBI, který jsme původně vyvinuli pouze pro potřeby digitalizace RBČL, pokrývá dostatečně mnohé z potřeb, které jsou na digitalizaci lístkových kartoték kladeny jak ze strany knihovníků, tak z pohledu programátorů. Jedná se o komplexní nástroj, bohatý na různé funkcionality spojené s administrací digitalizované lístkové kartotéky. Produkt dosud neměl instalační nástroj. Věřím, že mnou připravená a odladěná jednoduchá instalace zjednoduší práci s aplikací a odstraní nutnost podrobného studia dokumentace i programového kódu před prvním spuštěním.

Testy prokázaly, že projekt RETROBI je vhodný i pro větší kartotéky a odolá řadě běžných útoků.

Už dnes je systém užíván k digitalizaci lístkových kartoték jiných organizací, konkrétně Institutem umění – Divadelním ústavem, v přípravách je nasazení pro Kabinet pro klasická studia filosofického ústavu AV ČR, v. v. i. a zájem projevilo i Národní muzeum. Zjednodušená instalace nepochybně přiláká další zájemce.

Projekt má obrovský potenciál dalšího rozvoje a pokud se časem SQL nadstavbou podaří odstranit nerelačnost databáze, stane se ještě cennějším.

Literatura

[1] Lístkové katalogy ústavů AV ČR - Knihovna Akademie věd ČR. *Knihovna Akademie věd ČR* [online]. 2012 [cit. 2014-12-01]. Dostupné z: http://www.lib.cas.cz/online-katalogy/souborne-katalogy/listkove-katalogy-ustavu-av-cr

[2] Alphabetische Kataloge der Bibliotheken an der Philosophischen Fakultät der Universität zu Köln. *Philosophische Fakultät der Universität Köln* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://retro-philfak.ub.uni-koeln.de

[3] *Apache CouchDB* [online]. The Apache Software Foundation, 2014 [cit. 2014-12-01]. Dostupné z: http://couchdb.apache.org

[4] *The Apache HTTP Server Project* [online]. The Apache Software Foundation, 2014 [cit. 2014-12-01]. Dostupné z: https://httpd.apache.org/

[5] *Apache Tomcat* [online]. The Apache Software Foundation, 2014 [cit. 2014-12-01]. Dostupné z: http://tomcat.apache.org

[6] *Apache Wicket* [online]. The Apache Software Foundation, 2014 [cit. 2014-12-01]. Dostupné z: https://wicket.apache.org/

[7] Domů - Staročeská sbírka [lístkový katalog]. *ÚČL AV ČR, v. v. i.* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://starocech.ucl.cas.cz

[8] Elektronischer Zettelkatalog AK I. *Universitätsbibliothek Greifswald* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://gso.gbv.de/DB=1.89/SET=19/TTL=1/NXT?FRST=1&SET=19

[9] excerpce - ABC.cz: slovník cizích slov. *ABC Českého Hospodářství a.s.* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://slovnik-cizich-slov.abz.cz/web.php/slovo/excerpce

[10] KATIF. *Národní knihovna* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://katif.nkp.cz/Katalogy.aspx?katkey=050BIBL19

[11] KatZoom Catalogs. *Österreichische Nationalbibliothek* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://katzoom.onb.ac.at/katzoom/libraries.htm

[12] Knihovní systém Clavius a LANius. *Lanius s.r.o.* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://www.lanius.cz

[13] KP-SYS - KpWin SQL. *KP-SYS spol. s r.o.* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://www.kpsys.cz

[14] *The leading OS for PC, tablet, phone and coud | Ubuntu* [online]. Canonical Ltd., 2014 [cit. 2014-12-01]. Dostupné z: http://www.ubuntu.com

[15] Lístkové katalogy | Moravská zemská knihovna v Brně. *Moravská zemská knihovna v Brně* [online]. 2014 [cit. 2014-12-01]. Dostupné z: https://listky.mzk.cz

[16] Lístkový katalog/ Card Catalogue - [KRY]. *Národní památkový ústav* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://www.npu.cz/html/302/knihovna/katalog/KRY/index.html

[17] Main Page - AppArmor. *AppArmor security project* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://wiki.apparmor.net

[18] *Main Page - SELinux Wiki* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://selinuxproject.org

[19] *ModSecurity: Open Source Web Application Firewall* [online]. Trustwave Holdings, Inc., 2014 [cit. 2014-12-01]. Dostupné z: https://www.modsecurity.org

[20] *Nikto2 | CIRT.net* [online]. CIRT.net, 2014 [cit. 2014-12-01]. Dostupné z: https://cirt.net/Nikto2

[21] O katalogu - nápověda. *Národní památkový ústav* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://www.npu.cz/html/302/knihovna/help.html

[22] Sdružení knihoven ČR. *Sdružení knihoven ČR* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://www.sdruk.cz/

[23] SLUB - Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden. *Staats- und Universitätsbibliothek Dresden* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://image.slub-dresden.de

[24] Technologie RETROKON - Projekt RETROKON. *Národní knihovna* [online]. Praha, 2014 [cit. 2014-12-01]. Dostupné z: http://retrokon.nkp.cz/retrokon

[25] Tritius. *Lanius s.r.o.* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://www.tritius.cz

[26] UCL Library Services Card Catalogue Online. *University College London* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://cardcat.ucl.ac.uk/cgi-bin/cardbrowse.pl?ctype=&drawer=4&card=1&max=944

[27] UCL Library Services Card Catalogue Online. *University College London* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://cardcat.ucl.ac.uk

[28] Universitätsbibliothek Wien : KatZoom : Bandkatalog bis 1931. *Universität Wien* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://www.univie.ac.at/ubwdb/cgi-bin/katzmf.cgi?katalog=1&faktor=3

[29] Universitätsbibliothek Wien: Zettelkatalog KATZOOM. *Universität Wien* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://www.univie.ac.at/ubwdb/cgi-bin/katzoom.cgi?katalog=1&faktor=8&shmode=2&tnr=7&frmnr=0

[30] VISK - Národní program retrospektivní konverze katalogů knihoven v ČR - RETROKOn. *Národní knihovna* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://visk.nkp.cz

[31] VKOL - GEK (1450-1500) - Seznam titulů. *Vědecká knihovna v Olomouci* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://aleph.vkol.cz

[32] Vyhledávání/ 1 Retrospektivní biliografie české literární vědy. *ÚČL AV ČR, v. v. i.* [online]. 2014 [cit. 2014-12-01]. Dostupné z: http://retrobi.ucl.cas.cz

[33] kolektiv autorů: Základní statistické údaje o kultuře v České republice 2011, I. Kulturní dědictví. *Národní informační a poradenské středisko pro kulturu* [online]. Praha, 2012 [cit. 2014-12-01]. Dostupné z: http://www.nipos-mk.cz/wp-content/uploads/2009/03/Statistika_kultury_2011_I.KULTURNI-DEDICTVI_web.pdf

[34] kolektiv autorů: Základní statistické údaje o kultuře v České republice 2011, III. Knihovny a vydavatelská činnost. *Národní informační a poradenské středisko pro kulturu* [online]. Praha, 2012 [cit. 2014-12-01]. Dostupné z: http://www.nipos-mk.cz/wp-content/uploads/2009/03/Statistika_kultury_2011_I.KULTURNI-DEDICTVI_web.pdf

[35] Hordějčuk, V.: *Retrospektivní bibliografie české literární vědy, Implementační dokumentace.* 2012.

[36] Malínek, V.: Digitalizace lístkové kartotéky Retrospektivní bibliografie české literatury. *IKAROS, elektronický časopis o informační společnosti*

[online]. Ikaros, 2012 [cit. 2014-12-01], ISSN 1212-5075. Dostupné z: http://ikaros.cz/node/7683

[37] Malínek, V.: Nápověda - Retrospektivní bibliografie české literatury. *ÚČL AV ČR, v. v. i.* [online]. Praha, 2014 [cit. 2014-12-01]. Dostupné z: http://retrobi.ucl.cas.cz/retrobi/napoveda.1

[38] Poskierová, L.: Retrokonverze. *WikiKnihovna* [online]. 2014-08-27 [cit. 2014-12-01]. Dostupné z: http://wiki.knihovna.cz/index.php/Retrokonverze

[39] Stoklasová, B.: NÁRODNÍ PROGRAM RETROSPEKTIVNÍ KONVERZE ČR. *Sdružení knihoven ČR* [online]. 1995 [cit. 2014-12-01]. Dostupné z: http://www.sdruk.cz/data/xinha/sdruk/ks1995/1995-3-204.pdf

Příloha **A**

Seznam použitých termínů

atribut Položka společná všem záznamům.

excerpce Vypisování, pořizování výpisků, výtahu z něčeho[9]. Podrobně popsáno v sekci 4.1, strana 29.

index Odkaz do kartotéky. Definuje hranici mezi skupinou záznamů.

kartotéka Ucelený fyzických soubor dat v listinné podobě určený k digitalizaci.

klíčový atribut Položka záznamu sloužící k třídění.

lístek Fyzický list papíru, tvořící jeden záznam nebo část záznamu.

podšuplík Umístění zarážek v rámci fyzického šuplíku, jedná se o dodatečné jemnější dělení záznamů podle primárního třídění. Podšuplík je vytvářen pro účely snazší adresace záznamu v šuplíku.

primární třídění Způsob, podle jakého klíče jsou záznamy v kartotéce tříděny (abecední, číselný apod.)

šuplík Skutečný fyzický šuplík obsahující jednotlivé záznamy kartotéky. Záznamy jsou v šuplíku řazeny zpravidla abecedně, případně podle jiného kritéria, jako je desetinné třídění nebo vzácněji podle jiných klíčů. Z důvodu přehlednosti a rychlého vyhledávání obsahuje šuplík záznamy pouze podle určité části klíče, to jest například celá písmena abecedy, případně rozdělení do jemnějších celků podle dalších písmen.

validovaný atribut Položka záznamu, jejíž hodnota je předem stanovena výčtem možných hodnot. Jedná se například o názvy měst a jiných geografických celků, názvy profesí, předmětů daného oboru a podobně.

záznam Jednotka obsahu kartotéky. V případě lístkových kartoték se jedná o jeden nebo více lístků obsahujících informace o daném předmětu.

Příloha **B**

Seznam použitých zkratek

ARL Advanced Rapid Library

DDoS Distributed Denial Of Service

DoS Denial Of Service

IPAC Internet Public Access Catalogue

IPS Intrusion Prevention System

MARC Machine Readable Cataloging

MŠMT Ministerstvo školství, mládeže a tělovýchovy

NIPOS Národní informační a poradenské střediska pro kulturu

NK ČR Národní knihovna české republiky

NPÚ Národní památkový ústav

NRIS Národní informační systém pro retrokonverzi

OPAC Online Public Access Catalogue

OWASP The Open Web Application Security Project

RBČL Retrospektivní bibliografie české literatury (obsah systému RETROBI)

SLI Středisko literárněvědných informací

ÚČL AV ČR v. v. i. Ústav pro českou literaturu Akademie věd České republiky, veřejná výzkumná instituce

VISK Veřejné informační služby knihoven

Příloha C

Obsah přiloženého CD

```
readme.txt ..................................stručný popis obsahu CD
retrobi.log
  modsec_audit.log ........ výpis logu simulovaného útoku na aplikaci
src
  retrobi.sh ........adresář s instalačními skripty aplikace RETROBI
  retrobisrc ........... adresář se zrojovými kódy aplikace RETROBI
text
  thesis.pdf ............................ text práce ve formátu PDF
  src-latex ..................adresář zdrojů práce ve formátu LaTeX
```